Data Quality

Cinzia Cappiello
cinzia.cappiello@polimi.it

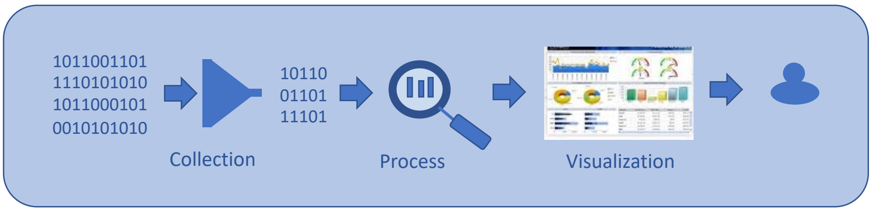
1

The importance of Data Quality

2

Data Driven Management

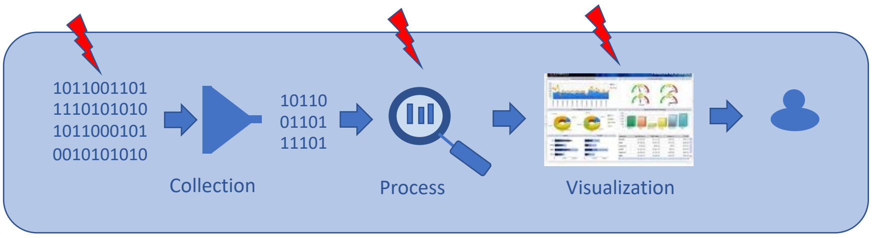
Data-driven Management is characterized by the practice of collecting data, analyzing it, and **basing decisions on** insights derived from the **information**.



https://www.smartsheet.com/data-driven-decision-making-management

3

The Problem: GIGO (Garbage In – Garbage Out) Phenomenon



The success of data-driven decision making depends on
- the **quality of data** collected
- the **methods** used to **analyze data**

4

even if it is a small error...you can have the snowball effect

5

Data Quality Horror stories

The Mars *Climate Orbiter,* a key part of NASA's program to explore the planet Mars, vanished in September 1999 after rockets were fired to bring it into orbit of the planet. It was later discovered by an investigative board that NASA engineers failed to convert English measures of rocket thrusts to newtons, a metric system measuring rocket force, and that was the root cause of the loss of the spacecraft. The orbiter smashed into the planet instead of reaching a safe orbit.

This discrepancy between the two measures, which was relatively small, caused the orbiter to approach Mars at too low an altitude. The result was the loss of a $125 million spacecraft and a significant setback in NASA's ability to explore Mars.

6

7

Data Quality Horror stories

Business

Why Britain has 17,000 pregnant men

By Sarah Kliff
April 7, 2012

https://www.washingtonpost.com/blogs/ezra-klein/post/why-britain-has-17000-pregnant-men/2012/04/06/gIQAC2oJ0S_blog.html

News > World > Americas

Workers demolish wrong house after relying on Google Maps for directions

Crew reportedly thought they had torn down the correct home - describing the situation as 'not a big deal'

Friday 25 March 2016 18:39 • Comments

https://www.independent.co.uk/news/world/americas/lindsay-diaz-google-maps-demolition-house-home-accident-a6952356.html

Spreadsheet error led to Edinburgh hospital opening delay

26 August 2020

https://www.bbc.com/news/uk-scotland-edinburgh-east-fife-53893101

30/10/23

7

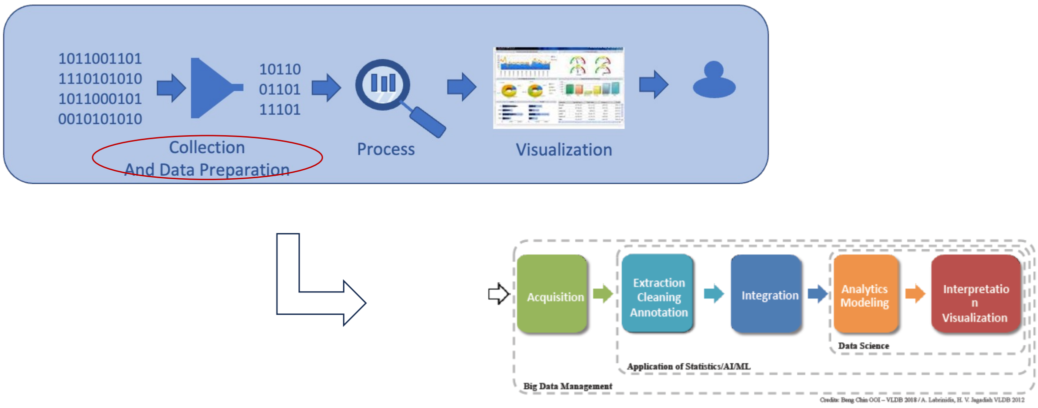
Data Quality Horror stories

Back in 1870," Arbesman explained, "Erich von Wolf, a German chemist, examined the amount of iron within spinach, among many other green vegetables. In recording his findings, von Wolf accidentally misplaced a decimal point when transcribing data from his notebook, changing the iron content in spinach by an order of magnitude. While there are actually only 3.5 milligrams of iron in a 100-gram serving of spinach, the accepted fact became 35 milligrams. Once this incorrect number was printed, spinach's nutritional value became legendary. So when Popeye was created, studio executives recommended he eat spinach for his strength, due to its vaunted health properties, and apparently Popeye helped increase American consumption of spinach by a third!"

http://www.ocdqblog.com/home/popeye-spinach-and-data-quality.html

8

We need an adequate architecture for analyze data



9

Why is data preparation important?

- Real-word data is often incomplete, inconsistent, and contain many errors…
- Data preparation, cleaning, and transformation comprises the majority of the work in a data mining application (90%).

10

Data Quality definition

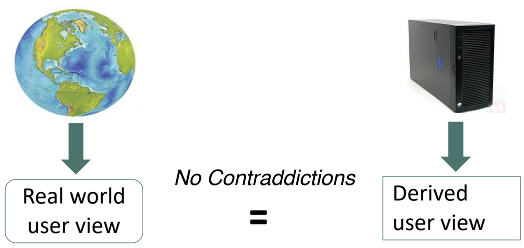
11

Data Quality definition

- Traditional definition

*“**Fitness for use** … the ability of a data collection to meet user requirements”*

- From an Information System perspective



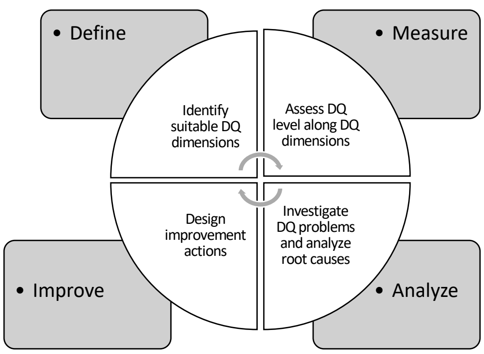
12

Data Quality Management

- Quality dimensions definition
- Quality dimensions assessment
- Quality issues analysis
- Quality improvement

13

Data Quality methodology



Wang R.Y., A Product Perspective on Total Data Quality Management, Communications of the ACM, Volume 41, Number 2, 1998

14

Data Quality dimensions

15

Data Quality issues to consider in data preparation activities are mainly related to…

- Missing values
- Duplicate data
- Inconsistent data
- Outliers
- Noise

16

Is Data Quality Measurable?

Data Quality
dimensions
&
Metrics

17

Data Quality Problems (single source) - example

ID	Name	Street and house number	Postcode	Town	Date of birth	Phone	e-mail
1	Janet Gordon	30 Fruit Street	75201	Dallas			
2	Kathy Robert	436 Devon Park Drive	94105	San Francisco	08.08.1969	215-367-2355	krob@robert.com
3	Sandra Powels	3349 North Ridge Avenue	33706	St. Pete Beach			
4	Johnstone, Jeffrey	3300 Sylvester Rd	92020	El Cajon			
5	Lowe Ruth-Hanna	25 Peachtree Lane	02112	Boston	10.10.50	(0617)-8845123	
6	Gordon Janet	30 Fruit Street	75201	Dallas			
7	NIck Goodman	Regional Campuses, 711	10020	New York	08/07/1975		n.good@goodman.com
8	Poweles Donna S.	3347 North Ridge	33706	Saint Pete Beach			
9	Cathy Robbert	436 Devon Park Drive	94105	San Francisco	08.03.1969		
10	Ruthanna Lowe	25 Peachtree Lane	02112	Boston		0617-8845123	
11	John Smith	10 Main Street	02112	New York			
12	Robert Katrin	434 Devon Park	94105	San Francisco			
13	Nick Goodman	56 Grafton Street	94105	San Francisco	08/07/1975		n.good@goodman.com
14	Sandro Powels	3349 North Ridge Av.	33706	Pete Beach			

18

Data Quality problems in BI (multiple sources) - example

ID	Diagnosis	Hospital	Province	Date	Cost
1	Flu	SR	Milan	01/05/2008	200
2	Flu	SR	Milan	24/5/2008	180-220
3	Flu	SR	Milan	04/05/2008	9999
4	Influenza	SC	Trento	03.05.2008	
5	Influenza	SC	Trento	03.04.2008	230
6	Influenza	SC	Trento	10.07.2008	
7	Flu Type A	CG	Milano	04-04-2008	130
8	Flu	OS	Bolzano	2008/04/23	130
9	Flu	OS	Bolzano	2008/05/11	200

19

Poor data quality is due mainly to

Missing values

Duplicates

Inconsistencies

Outliers

Noise

Out-of-date data

20

Most used objective Dimensions

Accuracy
- the extent to which data are correct, reliable and certified

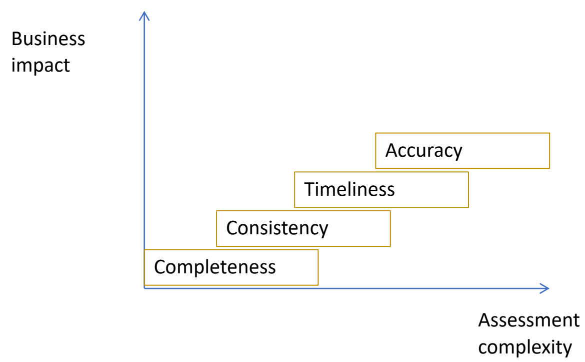
Completeness
- the degree to which a given data collection includes the data describing the corresponding set of real-world objects

Consistency
- the satisfaction of semantic rules defined over a set of data items

Timeliness
- the extent to which data are sufficiently up-to-date for a task

21

Assessment complexity



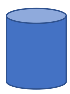
22

Data Quality Improvement

23

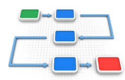
Data Quality improvement strategies

Data-based approaches



- They focus on data values and aim to identify and correct errors without considering the process and context in which they will be used

Process-based actions



- They are activated when an error occurs and aim to discover and eliminate the root cause of the error

24

Data Based approach: data cleaning

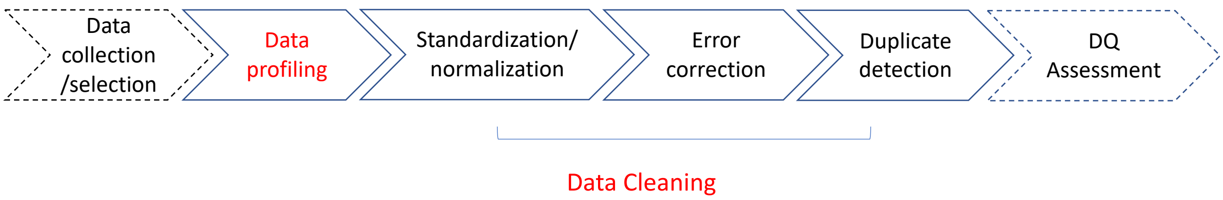
Definition

"Data cleaning is the process of identifying and eliminating inconsistencies, discrepancies and errors in data in order to improve quality"

[Naumann 2000]

25

Steps of Data Cleaning



Felix Naumann, Kai-Uwe Sattler 2006

26

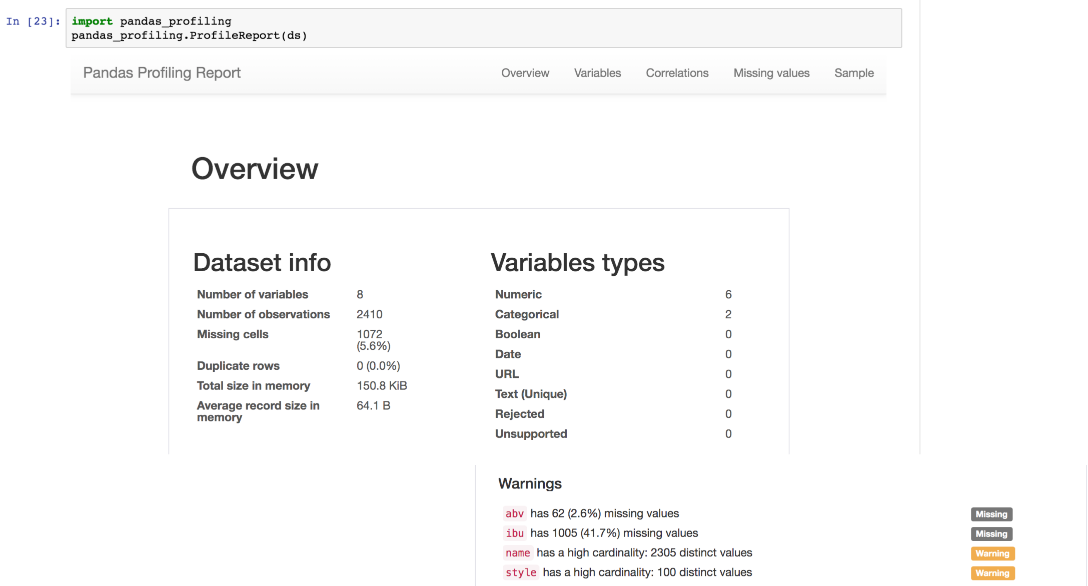
Profiling

- Analysis of content and structure of attributes: Data type, domain, data distribution and variance, occurence of null values, uniqueness, format (e.g., mm/dd/yyyy)
- Analysis of dependencies between attributes of a single relation: E.g., Functional dependencies, primary key candidates
- Analysis of overlapping attributes from different relations: Redundancies, foreign keys
- Number of missing values or wrong values
  - current vs.expected cardinality
  - frequency of null values, minimum / maximum, variance
- Duplicates
  - Number of tuples vs. Cardinality of attribute domain

Felix Naumann, Kai-Uwe Sattler 2006

27

Data Profiling with Python: Ydata profiling



28

Data Profiling with Python: Ydata profiling

Sample

First rows

	abv	brewery_id	ibu	id	name	ounces	style	u
0	0.050	408	NaN	1436	Pub Beer	12.0	American Pale Lager	0
1	0.066	177	NaN	2265	Devil's Cup	12.0	American Pale Ale (APA)	1
2	0.071	177	NaN	2264	Rise of the Phoenix	12.0	American IPA	2
3	0.090	177	NaN	2263	Sinister	12.0	American Double / Imperial IPA	3
4	0.075	177	NaN	2262	Sex and Candy	12.0	American IPA	4
5	0.077	177	NaN	2261	Black Exodus	12.0	Oatmeal Stout	5
6	0.045	177	NaN	2260	Lake Street Express	12.0	American Pale Ale (APA)	6
7	0.065	177	NaN	2259	Foreman	12.0	American Porter	7
8	0.055	177	NaN	2258	Jade	12.0	American Pale Ale (APA)	8
9	0.086	177	NaN	2131	Cone Crusher	12.0	American Double / Imperial IPA	9

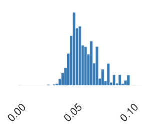
29

Data Profiling with Python: Ydata profiling

Variables

abv
Numeric

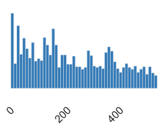
Distinct count	75	**Mean**	0.059773424
Unique (%)	3.1%	**Minimum**	0.001
Missing (%)	2.6%	**Maximum**	0.128
Missing (n)	62	**Zeros (%)**	0.0%
Infinite (%)	0.0%		
Infinite (n)	0		



Toggle details

brewery_id
Numeric

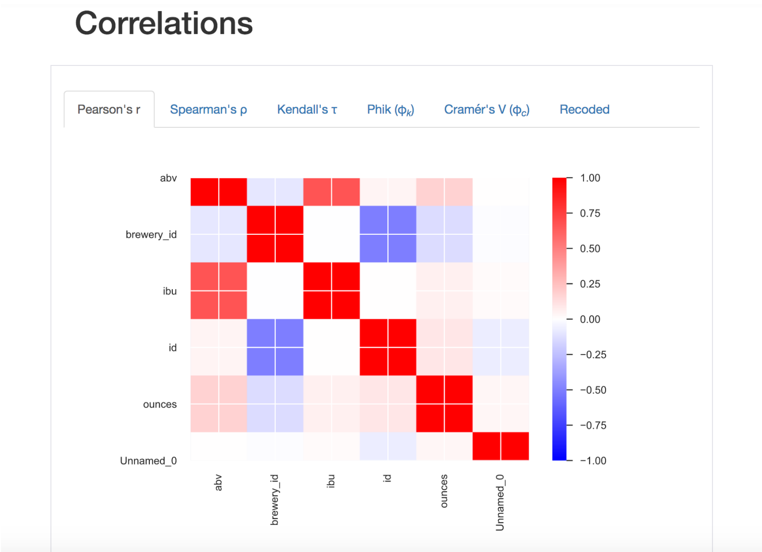
Distinct count	558	**Mean**	231.7497925
Unique (%)	23.2%	**Minimum**	0
Missing (%)	0.0%	**Maximum**	557
Missing (n)	0	**Zeros (%)**	0.2%
Infinite (%)	0.0%		
Infinite (n)	0		



Toggle details

30

Data Profiling with Python: Ydata profiling



31

Cleaning tasks

Normalization/standardization
- Datatype conversion
- Discretization
- Domain specific

Missing values
- Detection
- Imputing

Outlier detection
- Model
- Distance

Duplicate detection

32

Data transformation and normalization

Data type conversion: varchar → int

Normalization: mapping into a common format

- date: 03/01/15 → 01-MAR-2015
- currency: $ → €
- tokenizing: „Smith, Paul“→ „Smith“, „Paul“

Discretization of numerical values

Domain-specific transformations

- Surname, name→ Name surname
- St. → Street
- Address transformation using address databases
- Domain-specific product names/codes (e.g., in pharmacy)

Felix Naumann, Kai-Uwe Sattler 2006

33

Error Localization and correction

This activity can be seen as composed of:

- Localization and correction of inconsistencies
- Localize and correction of incomplete data
- Localization of outliers

34

Localize and correct inconsistencies

Once we have a valid, i.e., at least consistent, set of edits, we can use them to perform the activity of error localization.

In particular, we can check syntactic accuracy and inconsistencies

After the localization of erroneous records, in order to correct errors, we could perform the activity called *new data acquisition*

35

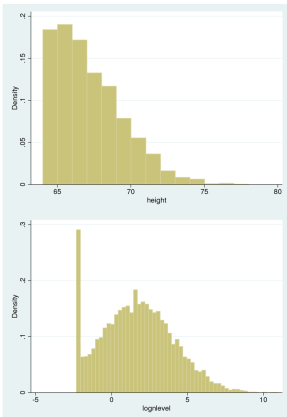
Missing data

Missing information on different levels

- Instance level: values, tuples, relation fragments, ...
- Schema level: Attributes, ...

Main Problems on instance level:

- Treating null values: missing value or default value?
- Data truncation and data censorization
- Biased data, e.g. caused by null values



36

Imputing missing value

unbiased estimators“
- Estimating missing values without changing characteristics of existing dataset (mean, variance, ...)
- E.g.: 1, 2, 3, _, 5 → (median: 2.75; variance: 4.659)

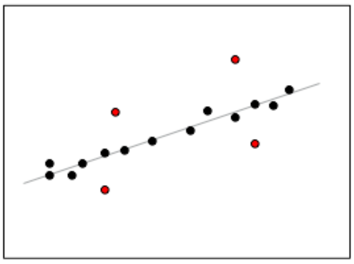
Exploiting functional dependencies
- E.g.: #Bedrooms → Income

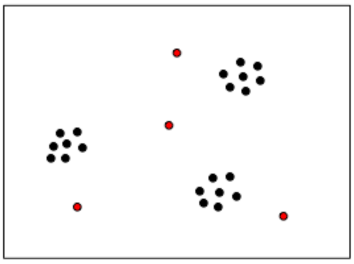
Techniques from statistics
- Linear regression:
  income = c • #Bedrooms
- techniques for non-linear dependencies:
  - Neural networks, ...

37

Outlier detection

Outlier: „suspicious“ observation that deviates too much from other observations. An outlier is then a value that is unusually larger or smaller in relation to other values in a set of data



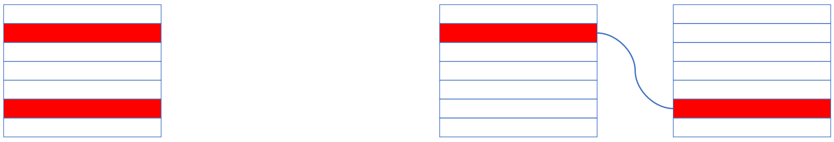


issues:
- detection: distribution, „geometry“, time series
- interpretation: data or observation error vs. real event

38

Duplicate detection Identify a good similarity measure

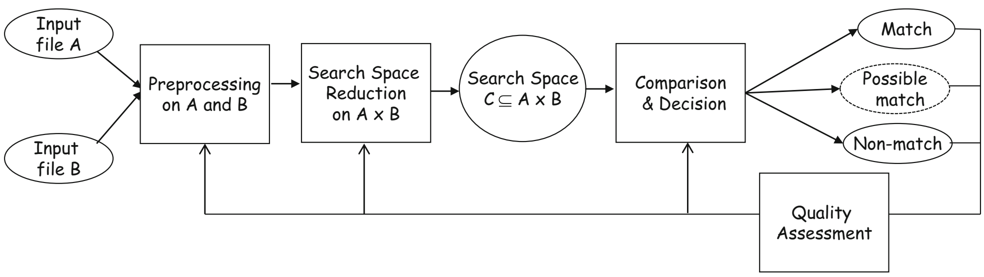
Duplicate detection (or entity reconciliation) is the discovery of multiple representations of the same real-world object.



- Main issues:
  - Identify a good similarity measure
  - Minimize the number of comparisons

39

The high level process



40

mailBatch for USA : Consumer-Comparison : Examples

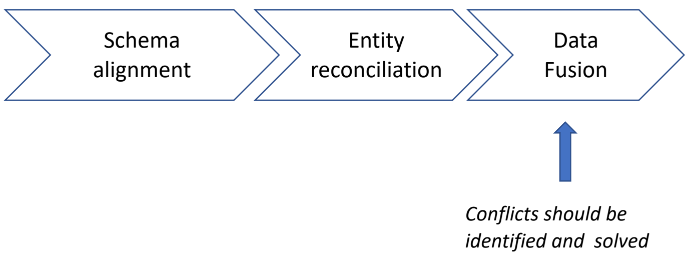
Clustering: Individual person level / Household level

Duplicates: Only certain / All

→	Sandra Powels	3349 North Ridge Avenue	33706	St. Pete Beach		
●	Powells Sandra	3349 North Ridge Avenue	33706	Pete Beach		
■	Poweles Donna S.	3347 North Ridge	33706	Saint Pete Beach		
→	Lowe Ruth-Hanna	25 Peachtree Lane	02114	Boston	10.10.50	(0617)-8845342
●	Ruthanna Lowe	1201 Oak Street	02132	Boston		0617-8845342
■	Lowe Ruth Anna		02110	Boston		
●	Ruth Lowe	1 Becton Drive	21030	Cockeysville	10.10.50	
→	Johnstone, Jeffrey	3300 Sylvester Rd	92020	El Cajon		
■	Jeffrey Johnstane	3300 Sylversterroad	92020	El Cajon		
●	J.R. Johnstone	3302 Sylvester	92020	Cajon		
■	Jeff Johnston	3300 S. Road	92020	El Cajon		
→	Gray-David Richard Crewson	Mail Stop, 300 Constitution Drive	33186	Miami		
■	Richard Crawson	300 constitution drive	33186	Miami		
●	Crewson, Gray Dave	Mail Stop, Constitution Dr. 301	33186	Miami		
■	Graham Crewsons	30 Constitution Drive	33186	Miami		
→	Michael & Nicole Goodman	Regionel Campuses, 711 Lincoln Bldg	10022	New York		
■	Ph. D. M. Goodnam	711 Lincoln Bldg	10022	New York		
●	Nicole Goodman	Regional Campuses, 711 Lincoln Bldg	10020	New York		
●	Michael Goodman	Regional Campuses, 711 Lincoln Bldg	10010	New York		
●	Mike Goodnan	711 Bldg	10020	New York		
→	Haddou, Judith Ben	137 Victoriacourt	22153	Springfield		
■	Benhaddou, Judith	137 S. Viktoria Court	22153	Springfield		
■	Haddou, Ben	137 Vicctoria Court	22153	Springfield		

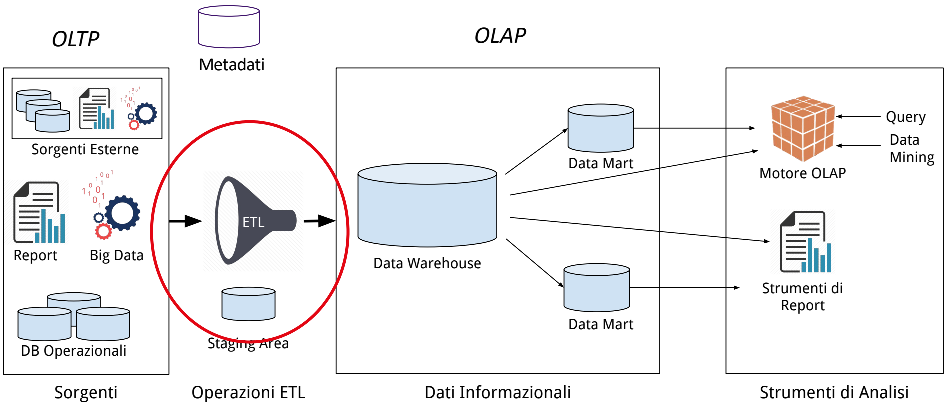
41

In case of multiple sources, data integration is also needed



42

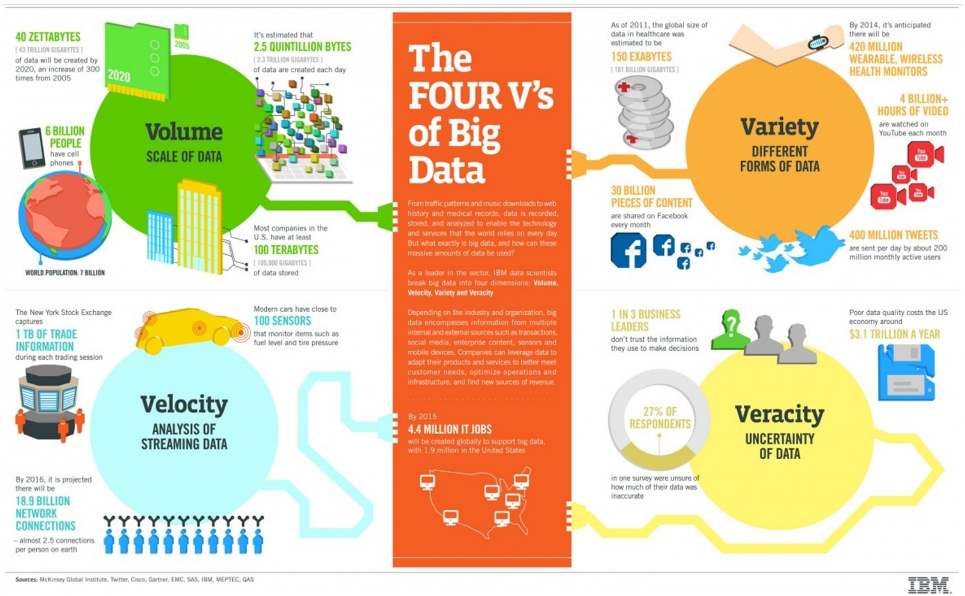
Data Quality improvement methods are also used in Data Warehouse



43

Big data and data quality

44



https://www.ibmbigdatahub.com/infographic/four-vs-big-data

45

Big data and data quality

Big Data analysis allows to understand customer needs, improve service quality, and predict and prevent risks.

High quality data are the precondition for guaranteeing the quality of the results of Big Data analysis.

Big Data tried to overcome **Data Quality issues with Data Quantity. But quality is still an issue.**

Cai, Li, and Yangyong Zhu. "The challenges of data quality and data quality assessment in the big data era." *Data Science Journal* 14 (2015).

46

Big data challenges

(1) Diversity of data sources (Variety)

Abundant data types - internal + external data sources

Complex data structures - structured, semi-structured, IoT

Difficult data integration - ETL and traditional approaches useless due to data volume and velocity

(2) Tremendous data volume (Volume)

Data quality profiling and assessment (collection, cleaning, and integration) is difficult to execute in a reasonable amount of time.

(3) Timeliness of data is very short (Velocity)

Data is updated continuously. If data is not collected and analysed in real time, information becomes outdated and invalid.

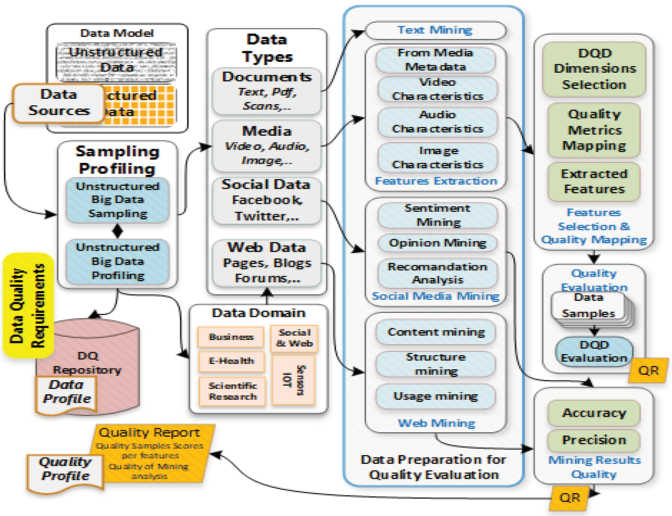
(4) Missing standard for Data Quality (Veracity)

Standards have been proposed for DQ of traditional data sources but not for big data.

Cai, Li, and Yangyong Zhu. "The challenges of data quality and data quality assessment in the big data era." *Data Science Journal* 14 (2015).

47

Unstructured Big Data Quality Assessment Model



I. Taleb, M. A. Serhani and R. Dssouli, "Big Data Quality Assessment Model for Unstructured Data," *2018 International Conference on Innovations in Information Technology (IIT)*, 2018, pp. 69-74, doi: 10.1109/INNOVATIONS.2018.8605945.

48

To summarize: most common DQ issues in big data

Not integrated data

Incomplete data

Incorrect data

Data cleaning have to be frequently performed

Inconsistent sources and issues in data integration

Source reliability

Data variety

Human resources: find the right competencies

Data provenance and lineage informatio should be available

49

50