Data Analysis Fundamentals

Lecture 4: Data Modeling

Vikas Kumar
Indian Institute of Technology Bombay

08th September 2023

Outline

- Understanding keys, cardinality and ER Diagrams
- Introduction to Data Quality
- The six dimensions of data quality
- Big Data

Unique keys

Features	**Primary keys**	**Unique keys**
Number of keys	One primary key in a parent table	One or more than one, in parent or child tables
Values	Must have a value, cannot be NULL	Can be a NULL value
Use	Identify every item in a table	Identify items in a table when they cannot have duplicate values
Ease	Cannot be removed, difficult to change	Can be removed or changed easily
Indexes	Clustered index	Non-clustered index

Unique keys

Student

Roll_no	Name	Class	Phone_no	Registration_no
1	Andrew	5	9854672256	895
2	Andrew	6	9955512456	564
3	Augosto	5		567

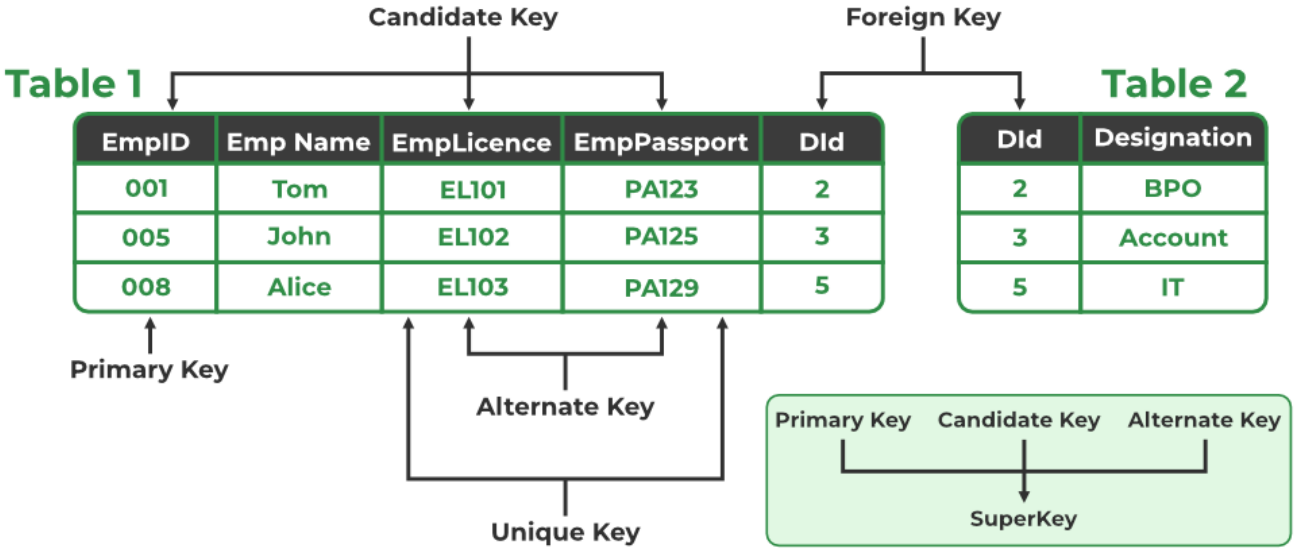
Primary key

Unique key

Unique key

Foreign Keys

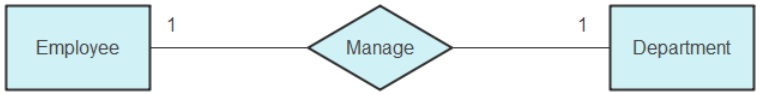
- The FOREIGN KEY constraint is a key used to link two tables together.
- A FOREIGN KEY is a field (or collection of fields) in one table that refers to the PRIMARY KEY in another table.



Cardinality

- Cardinality means how the entities are arranged to each other.
- The relationship structure between entities in a relationship set.

one-to-one (1:1)

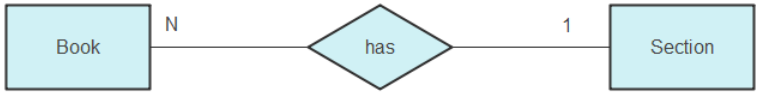


✓ One student can have only one student ID

one-to-many (1:N)

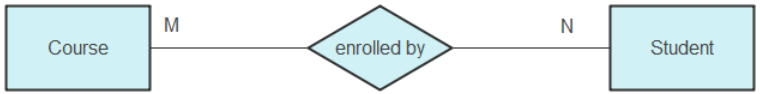
✓ Edyoda has multiple trainers

many-to-one (N:1)



✓ One trainer teaches multiple students in the class

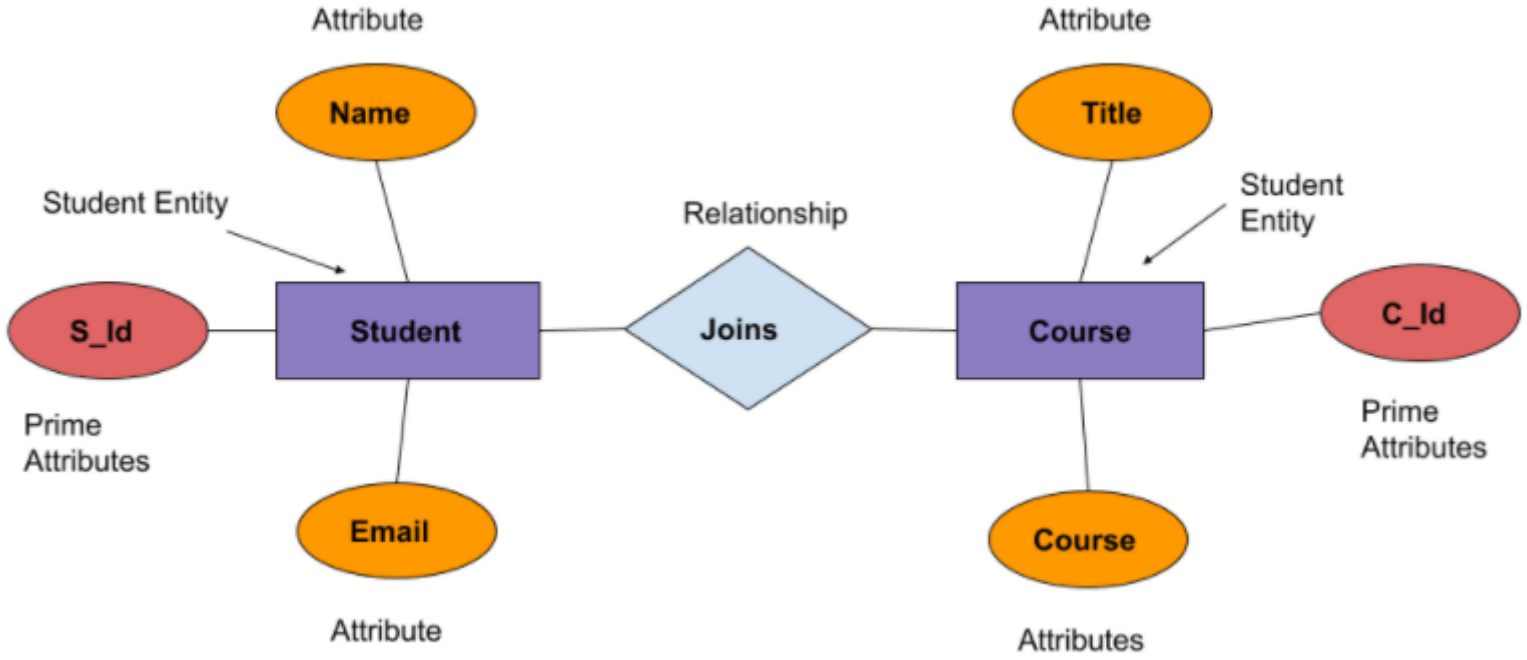
many-to-many (M:N)



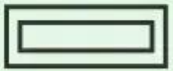
✓ One trainer takes several modules, and same module is taken by different trainers

Entity Relationship (ER) Diagram

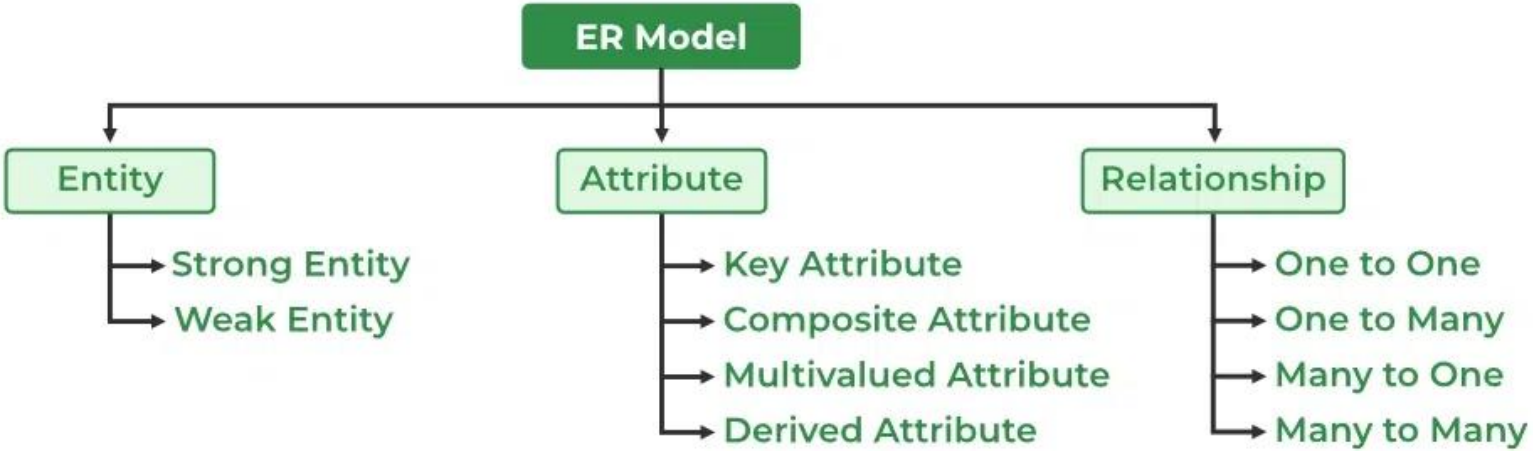
- A type of flowchart that illustrates how “entities” such as people, objects or concepts relate to each other within a system.



Entity Relationship (ER) Diagram

Figures	Symbols	Represents
Rectangle		Entities in ER Model
Ellipse		Attributes in ER Model
Diamond		Relationships among Entities
Line		Attributes to Entities and Entity Sets with Other Relationship Types
Double Ellipse		Multi-Valued Attributes
Double Rectangle		Weak Entity

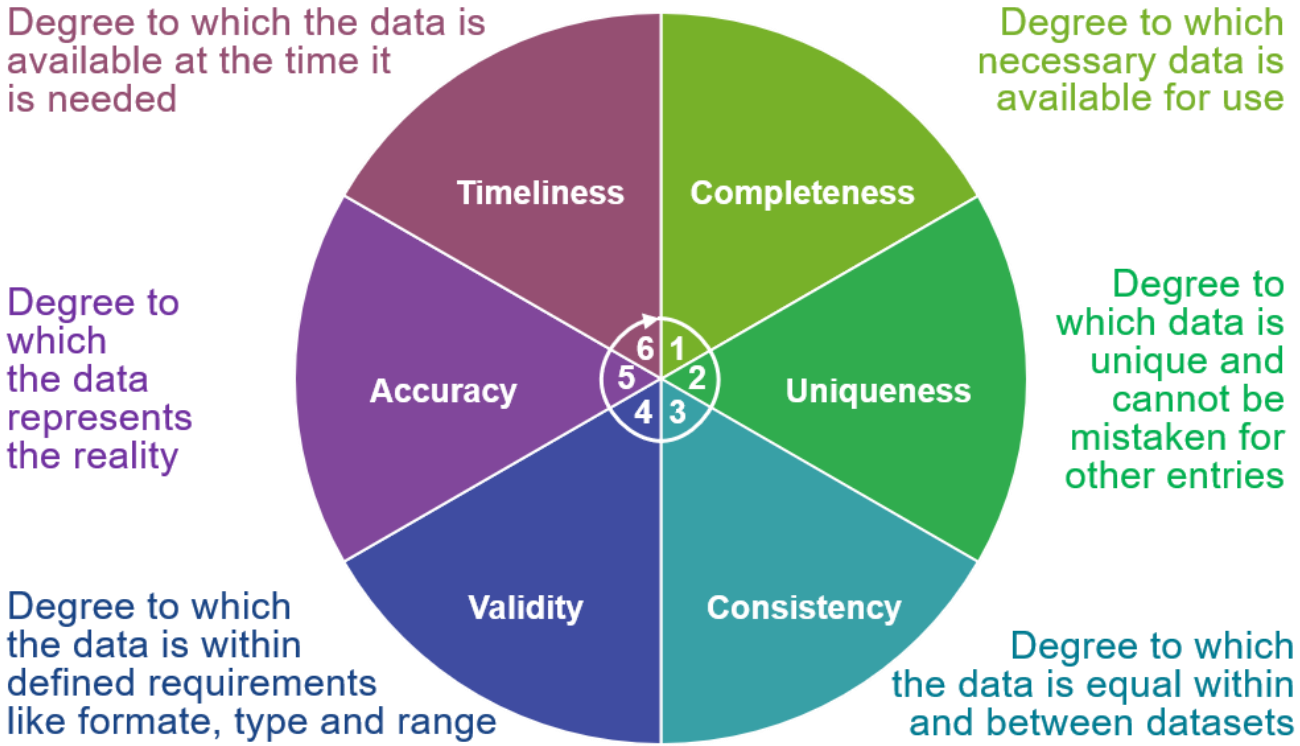
Entity Relationship (ER) Diagram



Data Quality

- ✓ Data quality measures how well-suited a data set is to serve its specific purpose.
- ✓ Data's suitability for a user's defined purpose.
- ✓ It is subjective, as the concept of quality is relative to the standards defined by the end users' expectations.

Six Dimensions of Data Quality



Completeness

- ✓ Data is considered “complete” when it fulfils expectations of comprehensiveness.
- ✓ The completeness data quality dimension is defined as the percentage of data populated vs. the possibility of 100% fulfilment.
- ✓ Missing Records Example: You are an eligible voter, but the record with your name is missing from the voter’s list at the voting booth.
- ✓ Null Attribute Example: Each customer record must have a name, email address, and phone. However, the phone number or the email ID might be missing in some of the customer records.

Completeness

Name	Email	Phone	Spend	Visit count	Reward points
Hank Williams	hank@msn.com	(564)342-1212	$210.03	2	47.14
Joe Panik		1415)321-7689	$37.45	1	35.00
David R Simcoke	devid@gmail.com		$59.13	2	30.00
R Kelly	rkelly@aol.com	(310)789-0000	$24.64	2	27.28
Bruce Bocily	bbochy@stgiants.com	(415)456-7890	$0.00	0	26.79
Buster Posey	buster@orgiants.com		$261.20	12	26.06
Klay Thompson	splashbro@Pwarriomerun	(510)543-2345	$0.00	0	25.24
Steve Kerr			$268.53	13	24.25
Ayeesha Curry	acurry@gmail.com	(510)426-7457	$407.52	8	19.70
Tim Lincecum	timmy@sfgiants.com	(415)453-2345	$3.272.99	126	19.23
P Diddy	diddy@outlook.com	(510)765-6789	$0.00	0	17.95

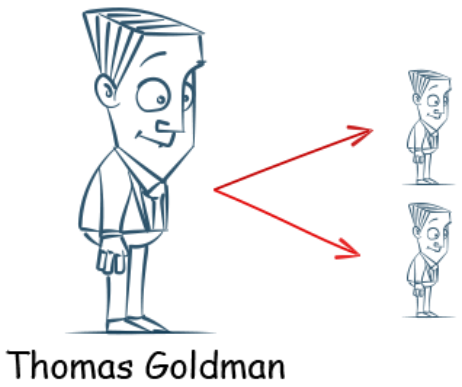
- ✓ Missing Reference Data
- ✓ Data Truncations

Validity

- Data validity describes the closeness of data value to predetermined values or a calculation.
- Signifies that the value attributes are available for aligning with the specific domain or requirement.
- For example, ZIP codes are valid if they contain the correct characters for the region.
- In a calendar, months are valid if they match the standard global names.
- Using business rules is a systematic approach to assess data validity.
- Any invalid data will affect the completeness of data.

Uniqueness

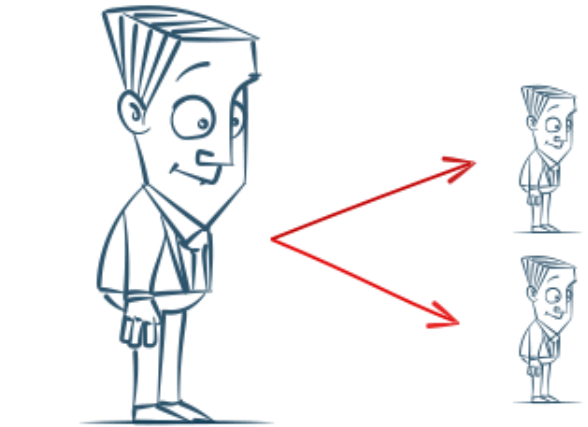
- The occurrence of an object or an event gets recorded multiple times in a dataset.
- An event or entity should only be recorded only once.
- Same Entity Is Represented With Different Identities



Customer Name	
Thomas Goldman	
Tom Goldman	

Uniqueness

- Same Entity Is Represented Multiple Times With Same Identity

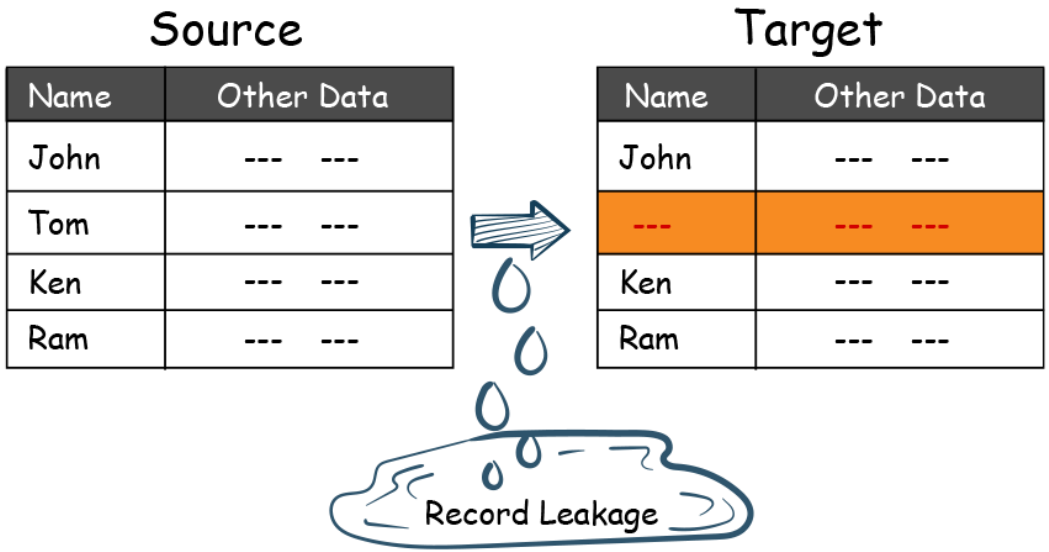


Customer Name	
Thomas Goldman	
Thomas Goldman	

Consistency

- Consistent data can be explained as how close your data aligns or is uniform with another reference dataset.

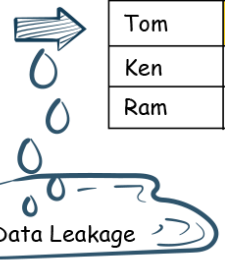
- Record Level Data Consistency Across Source and Target



Consistency

- Attribute Consistency Across Source And Target

Name	Email id	Phone no
John	jo@abc.com	203123425
Tom	tom@pqr.com	2129495555
Ken	ken@xyz.com	7181234219
Ram	ram@ghi.com	7772221234



Name	Email id	Phone no
John	jo@abc.com	203123425
Tom	---	2129495555
Ken	ken@xyz.com	---
Ram	ram@ghi.com	7772221234

- Attribute Consistency Across Source And Target

Order

Order ID	Item	Qty	$ Amt
0123	Gown	1	$150
0123	Dress Pant	3	$200

Shipment

Ship ID	Order ID	Item	Qty	Ship Dt
SHAB1	0123	Gown	3	2/2/2020
SHAB1	0123	Dress Pant	1	2/2/2020

Consistency

- Consistency In Data Representation Across Systems

1	Male
2	Female
3	Unknown

1	M
2	F
3	Uk

1	Male
2	Female

1	Male
2	Female
3	Unknown
4	NA

1	Alpha male
2	Male
3	Alpha female
4	Female
5	Unknown

- Transaction Data Consistency

Yesterday's Account Balance

Act	Dt	$ Amt
A 99	1/1/20	$ 4000
A 500	2/1/20	$ 9000

+

Today's Transaction

Act	Dt	Txn Amt
A 99	2/2/20	+ 1000
A 500	2/2/20	- 1000

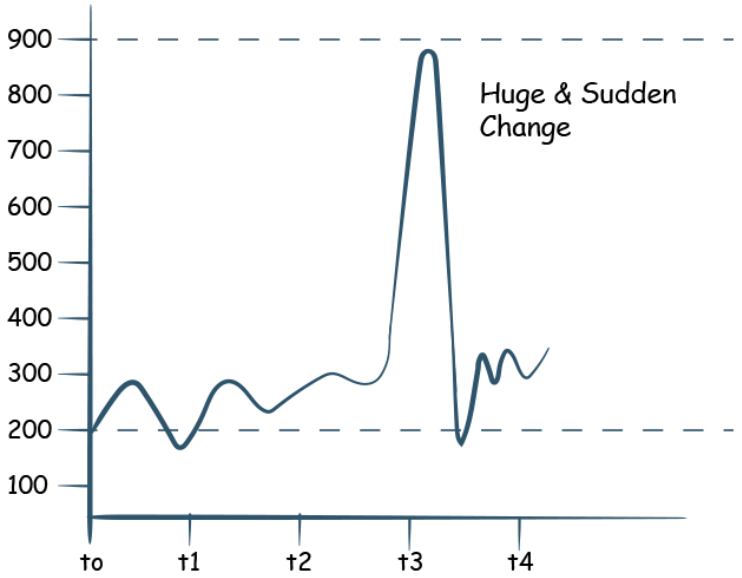
?
=

Today's Account Balance

Act	Dt	$ Amt
A 99	2/2/20	$ 5000
A 500	2/2/20	$ 4000

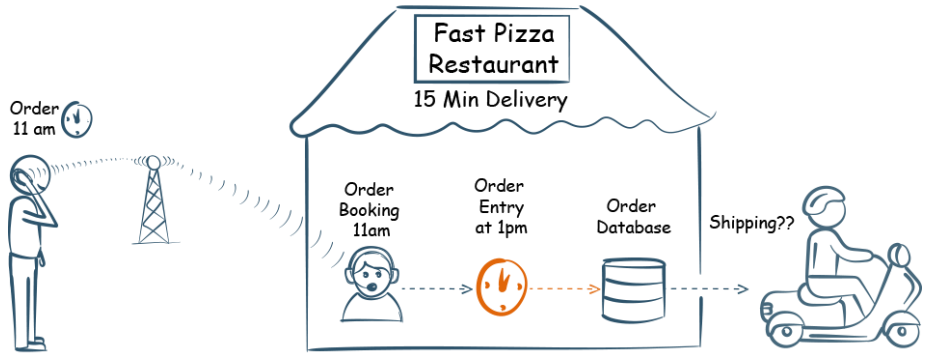
Consistency

- Data Consistency Over Time



Timeliness

- It is the time lag between the actual event time vs. the event captured in a system to make it available for use.
- The delay between actual event occurrence and the data availability exceptions by the business or the downstream process defines the timeliness quality dimension.
- It is important to understand that the data is still valid, just late.
- In self-driving cars, any lag in the arrival of data can cause accidents as it won't be able to course correct.



Accuracy

- The term “accuracy” refers to the degree to which information accurately reflects an event or object described.
- For example, if a customer’s age is 32, but the system says she’s 34, that information is inaccurate.
- Data accuracy is the degree to which data represent real-world things, events, or an agreed-upon source.

CPI Index Value

Month	Val CPI
2020 Jan	257.971
2020 Feb	258.678
2020 Mar	258.115

Table 1. Consumer Price Index for All Urban Consumers (CPI-U): U.S. March 2020
[1982-84=100, unless otherwise noted]

Expenditure category	Relative importance Feb. 2020	Unadjusted indexes		
		Mar. 2019	Feb. 2020	Mar. 2020
All items	100.000	254.202	258.678	258.115

Currency

- Data Currency reflects the real-world state vs. the state captured in the dataset.

- A mailing list has customers' addresses. But if the customers have already moved to a new address, the data loses its currency.

- Timeliness is the late arrival of data or delay, but the information is still accurate. If the data is late and reflects a state that has changed or expired, and hence the data becomes irrelevant and loses its value or currency.

Conformity

- Conformity means that the data values of the same attributes must be represented in a uniform format and data types.

- Format Conformity

Customer	Order Dt	---
Don	2019/2/12	--- ---
Joe	12/24/2019	--- ---
Tim	2019/2/12 22:09:01	--- ---

Conformity

- Data Type Conformity

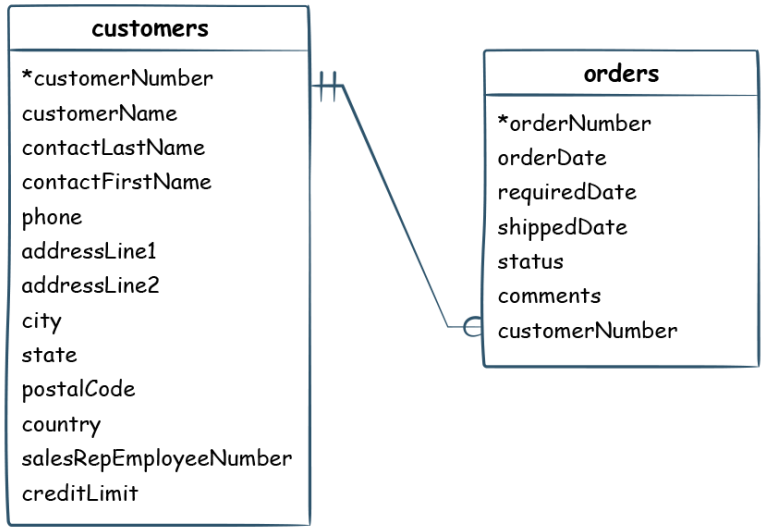
Customer	Order Amt	---
Don	$100	--- ---
Joe	Five Hundred Dollars	--- ---
Tim	$300	--- ---

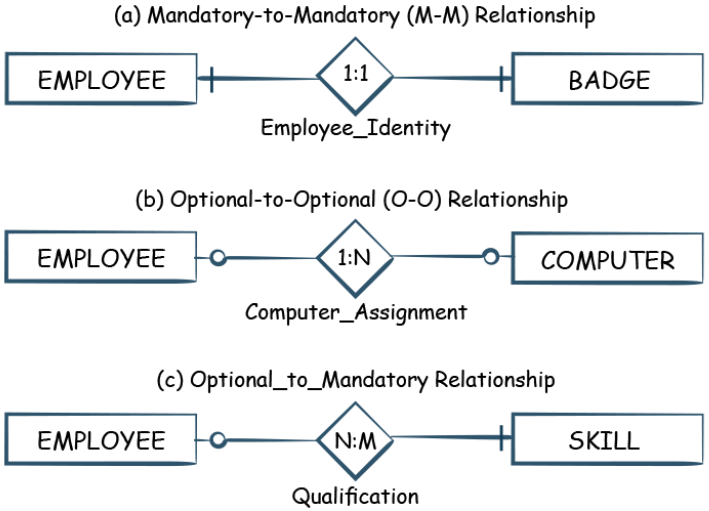
Integrity

- Data Integrity Quality dimension is the degree to which a defined relational constraint is implemented between two data sets.

- Referential Integrity Or Foreign Keys

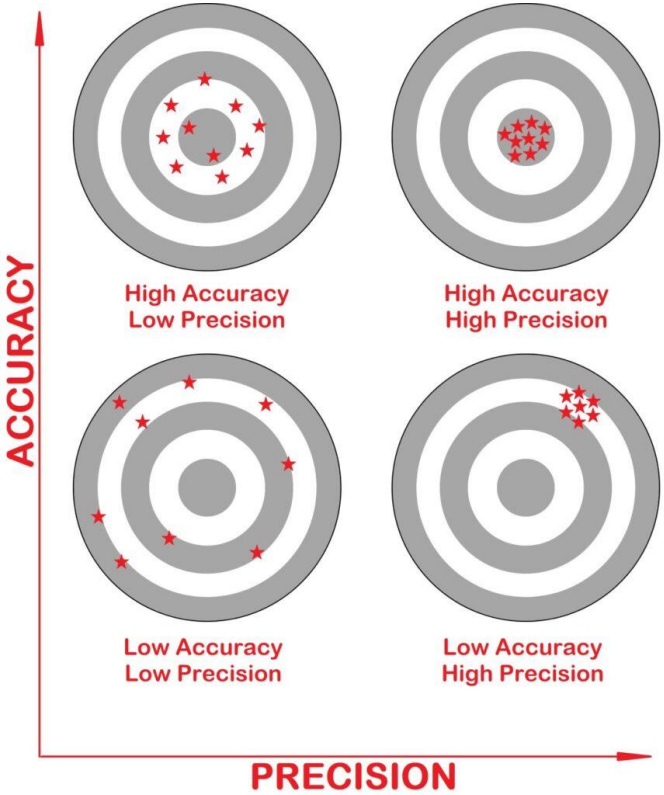
- Cardinality Integrity





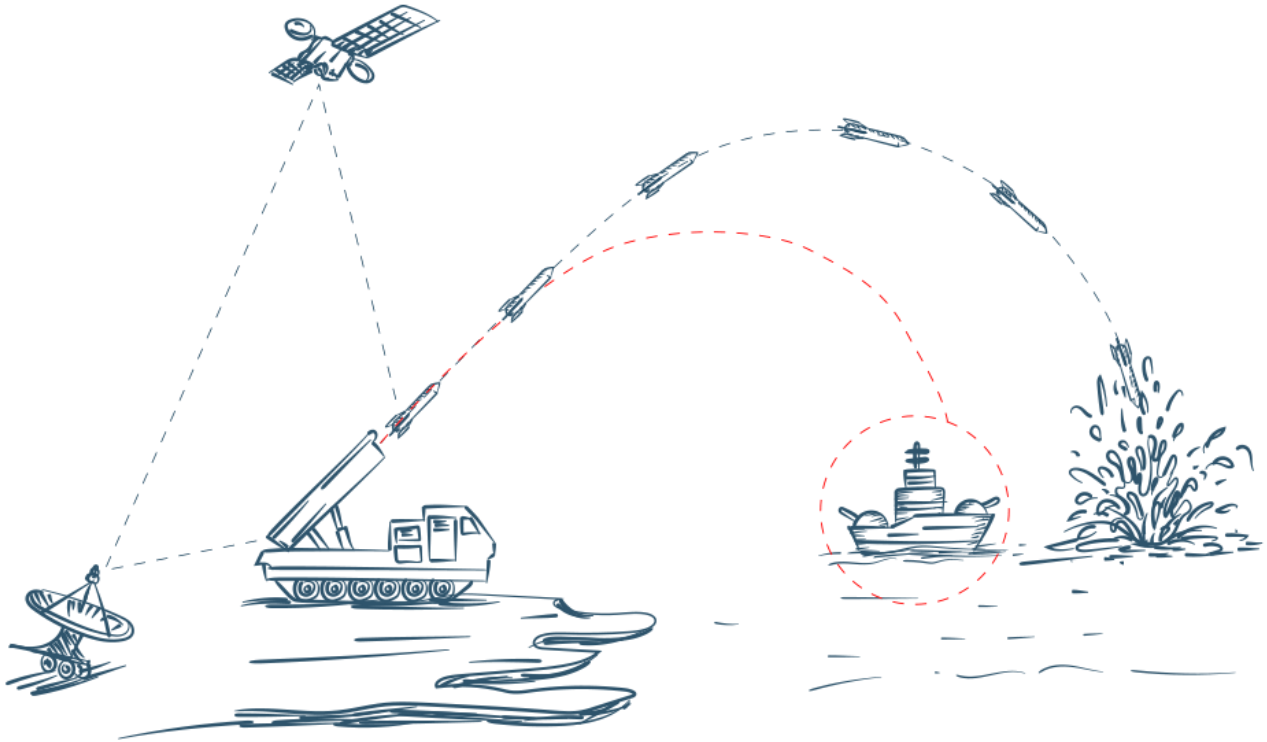
Precision

- The degree to which the data has been rounded or aggregated.



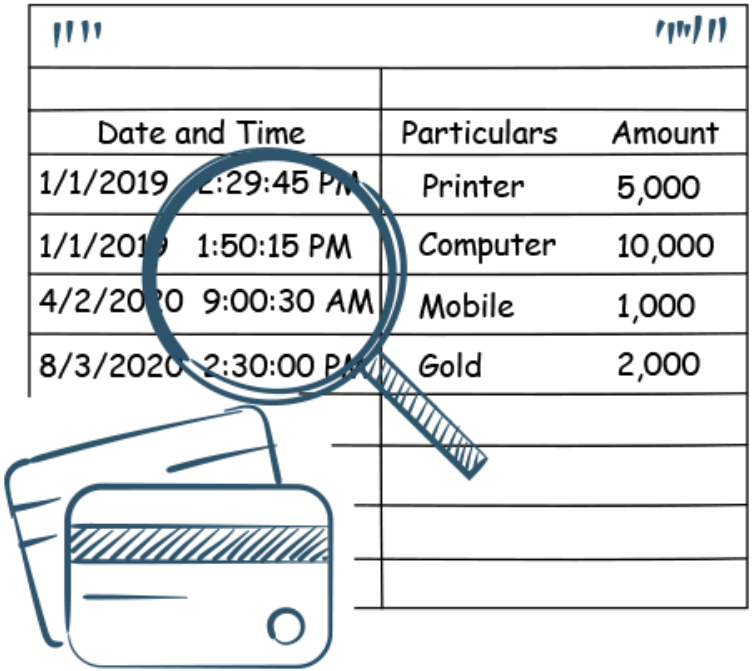
Precision

- Precision Errors Due To Rounding of Number



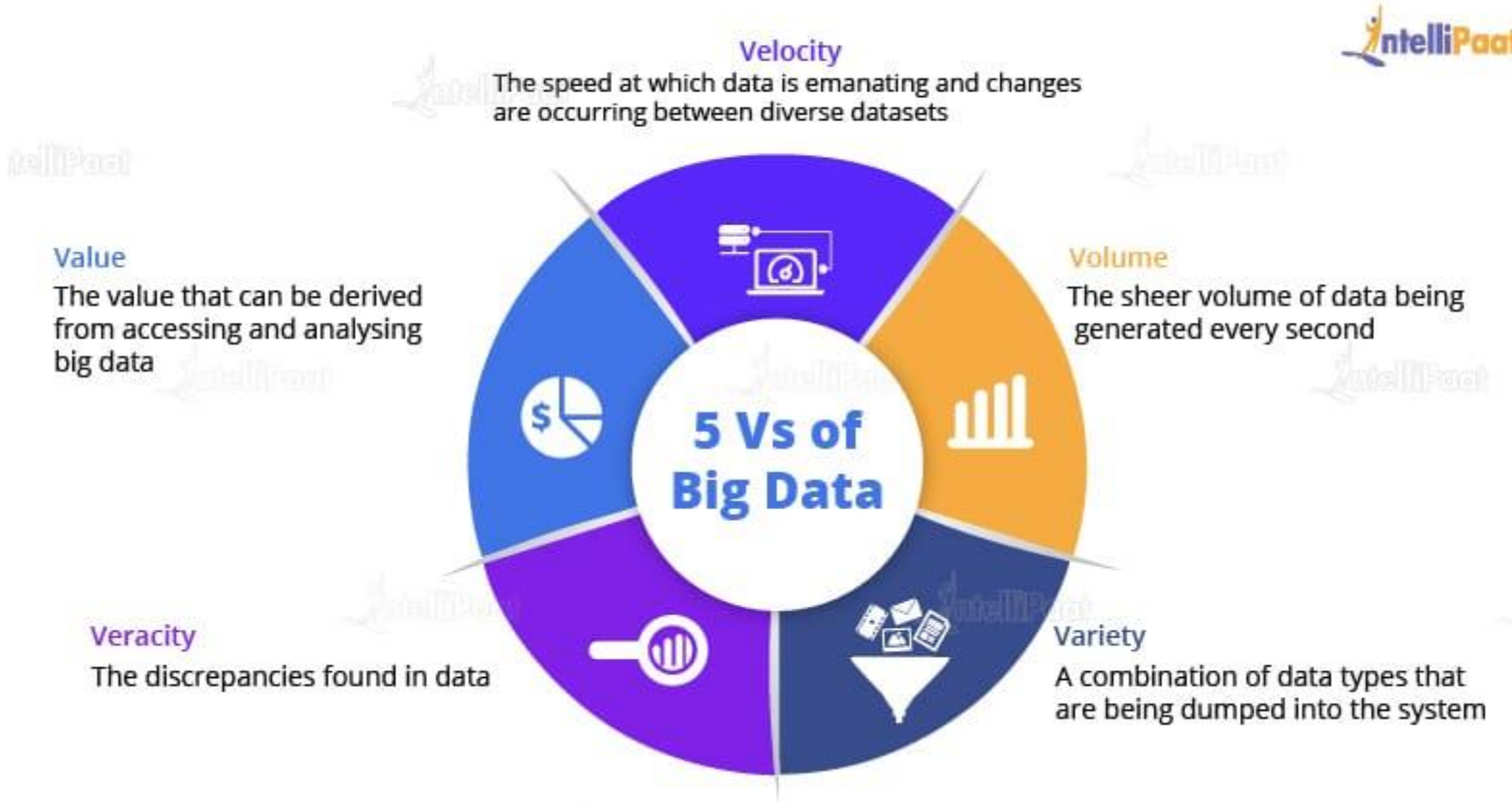
Precision

- Time Precision

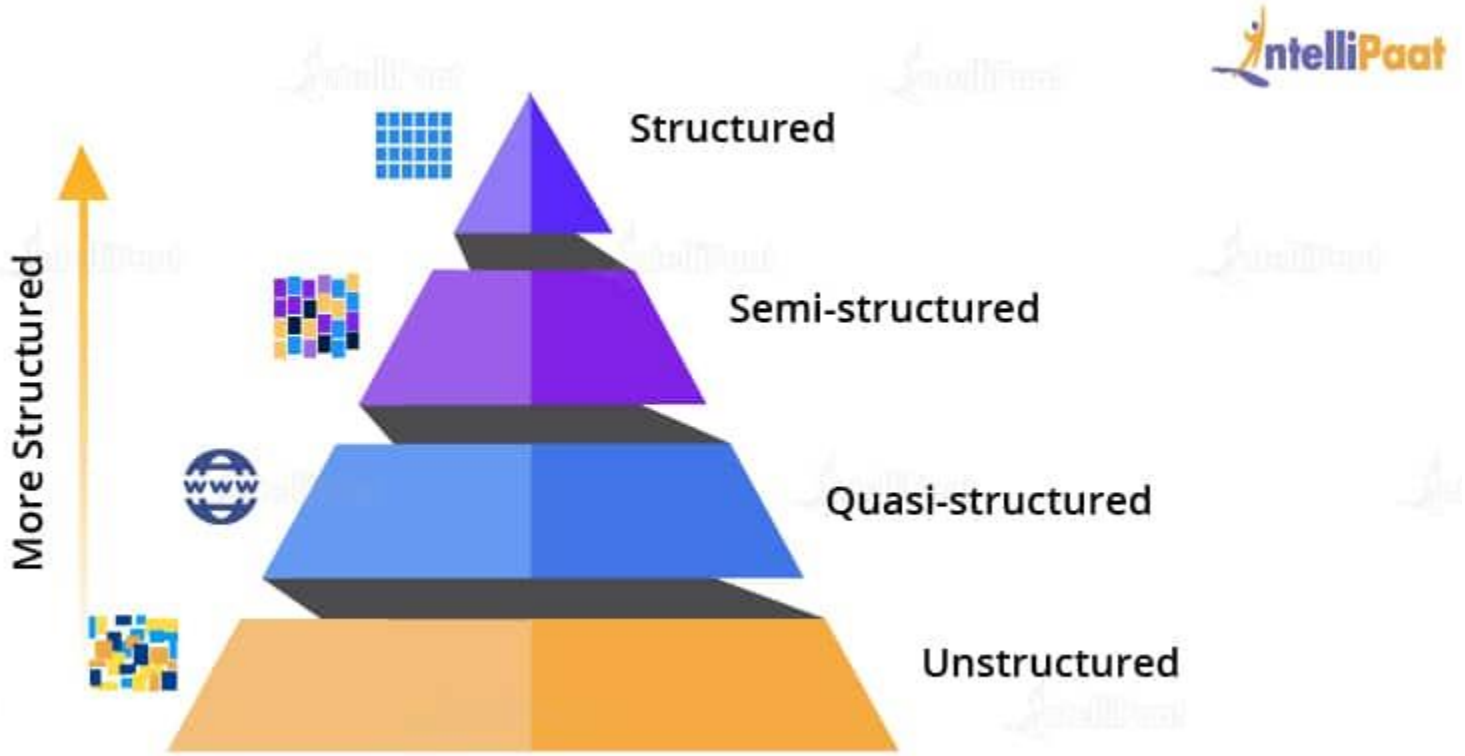


		Debit
Date	Particulars	Amount ($)
1/1/2019	Printer	5,000
1/1/2019	Computer	10,000
4/2/2020	Mobile	1,000
8/3/2020	Gold	2,000

Big Data

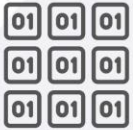

Velocity
The speed at which data is emanating and changes are occurring between diverse datasets
Value
The value that can be derived from accessing and analysing big data
Volume
The sheer volume of data being generated every second
5 Vs of Big Data
Veracity
The discrepancies found in data
Variety
A combination of data types that are being dumped into the system

Big Data


IntelliPaat
Structured
Semi-structured
Quasi-structured
Unstructured
More Structured
www

Big Data

Structured data



Characteristics

Predefined data models
Easy to search
Text-based
Shows what's happening

Resides in

Relational databases
Data warehouses

Stored in

Rows and columns

Examples

Dates, phone numbers, social security numbers, customer names, transaction info

Unstructured data

Characteristics

No predefined data models
Difficult to search
Text, pdf, images, video
Shows the why

Resides in

Applications
Data warehouses and lakes

Stored in

Various forms

Examples

Documents, emails and messages, conversation transcripts, image files, open-ended survey answers

Semi-structured data

Characteristics

Loosely organized
Meta-level structure that can contain unstructured data
HTML, XML, JSON

Resides in

Relational databases
Tagged-text format

Stored in

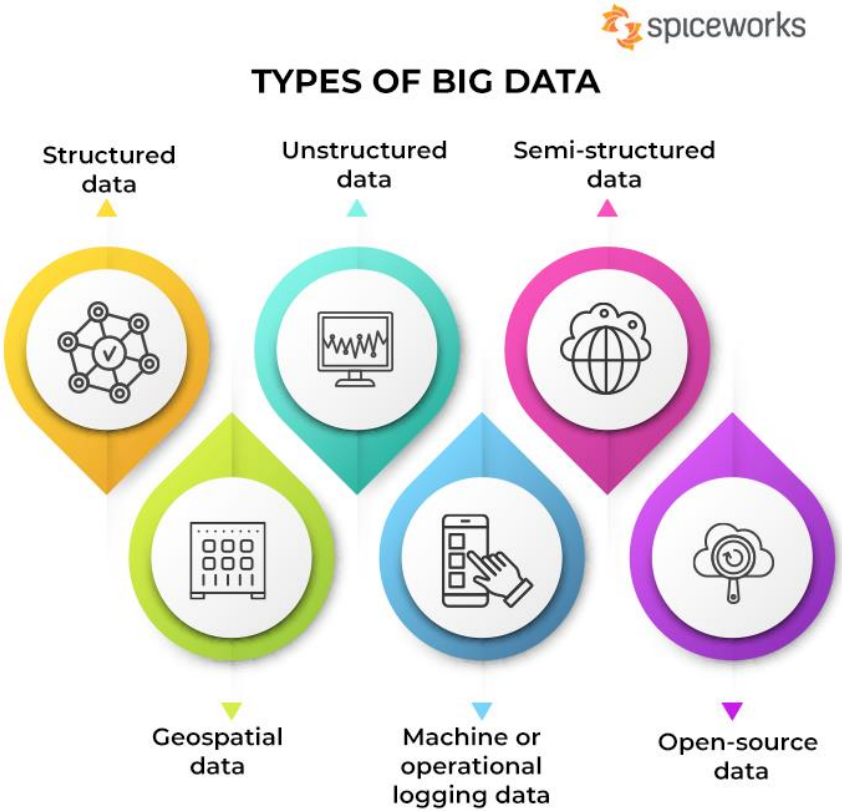
Abstracts & figures

Examples

Server logs, tweets organized by hashtags, emails sorting by folders (inbox; sent; draft)

LEVITY

Big Data


spiceworks
TYPES OF BIG DATA
Structured data
Unstructured data
Semi-structured data
Geospatial data
Machine or operational logging data
Open-source data


شكراً جزيلاً
ngiyabonga
рахмат
danke
謝謝
teşekkür ederim
spas
Баярлалаа
mersi
kia ora
barka
welalin
tack
mahalo
спасибо
faafetai lava
vinaka
blagodaram
dank je
misaotra
paldies
grazzi
tapadh leat
спасибі
matondo
kiitos
dankie
thank you
gracias
хвала
nanni
nandri
dhanyavad
mauruuru
köszönöm
aciu
asante
manana
bedankt
bayarlalaa
gràcie
hvala
dankon
obrigada
murakoze
tenki
enkosi
akun
djiere dieuf
tau
mochchakkeram
дякую
mamnun
dziękuję
chnorakaloutioun
gratias ago
gràcies
sulpáy
go raibh maith agat
chokrane
sobodi
dėkuji
sukriya
kop khun krap
taiku
obrigado
sagolun
arigatô
takk
dakujem
trugarez
mèsi
didi madloba
ありがとう
tanemirt
grazie
najis tuke
terima kasih
rahmet
ευχαριστώ
diolch
dhanyavadagalu
shukriya
mercé
мерси
kam sah hamnida
rahmat
তোমাকে ধন্যবাদ
감사합니다
xièxie
merci