Data Analysis Exercise

Transformation is necessary in the ETL process.

Identify the categories of data anomalies and suggest solutions on how to deal with them.

Student table

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Name | Address | Gender | location | Personal tutor |
| Aidan Andrews | 1 The Avenue, 1AA AAA | M | H100 | SL |
| Billy Brag | 2 Bank Street, 2BB BBB |  | CQ100 | CA |
| Colin Chips | 3 Crescent mews, 3CC CCC | M | DL |  |
| Denise Donnington | 4 The Drive, 4DD DDD | F | HK100 | SL |

Student Location table

|  |  |
| --- | --- |
| Location id | Location name |
| H100 | Headingley Campus, Leeds |
| CQ100 | Civic Quarter, Leeds |
| DL | Distance Learner |
| HK100 | Hong Kong |

Student union table

|  |  |  |  |
| --- | --- | --- | --- |
| Name | Home Address | Gender | location |
| Andrews, A | 13 The Avenue, 1AA AAA | Male | H100 |
| Brag, B | 2 Bank Street, 2BB BBB |  | CQ100 |
| Chips, C | 3 Crescent mews, 3CC CCC | Male | DL |
| Donington, D. E | 4- 6 The Drive, 4DD DDD | Female | HK200 |

The data anomalies are:

1. Name and address problem

* No unique key
* Missing values
* Different formats for the names
* Different addresses for the same student
* Different spelling and names for same student

Ways to solve the name and address problem

* Create atomic values, (use a sequence for ids)
* Standardise the formats
* Verify the data accuracy
* Match with other records
* Document using metadata

1. Missing value problems

Missing values for the Personal tutor.

Ways to solve the name and address problem

* Ignore
* Wait
* Flag rows
* Replace using IFNULL(SLO)

1. Multiple encoding

The example shows 2 data bases each storing the gender.

Db 1: M,F

Db 2: Male, Female ------🡪 Warehouse M,F

Transformation should ensure that one encoding is adopted.

1. Referential Integrity Problem

|  |  |  |  |
| --- | --- | --- | --- |
| Donington, D. E | 4- 6 The Drive, 4DD DDD | Female | HK200 |

Has been assigned to a location which doesn’t exist

Solutions:

* constraints