UNIT–3 : BI – Data Preparation (PYQ Solution)

AKTU · Business Intelligence and Analytics (BCDS051) · Solutions by **Chirag Jain**, Ycotes

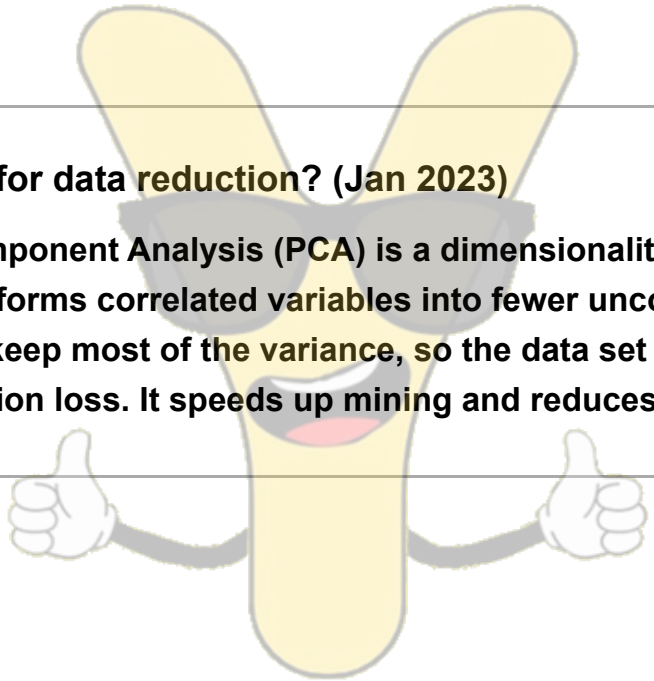
SECTION A Answers (≈70 Words Each)

Q1. Why is data validation used? (Jan 2023)

Data validation checks that data is accurate, complete, consistent and in the correct format before it is stored or analysed. It prevents wrong, duplicate or missing values from entering the system, protects the quality of reports and analytics, and reduces errors and rework, so decisions are based on reliable data.

Q2. Is PCA used for data reduction? (Jan 2023)

Yes. Principal Component Analysis (PCA) is a dimensionality reduction technique. It transforms correlated variables into fewer uncorrelated principal components that keep most of the variance, so the data set becomes smaller with little information loss. It speeds up mining and reduces noise.



SECTION B Answers (≈250 Words Each)

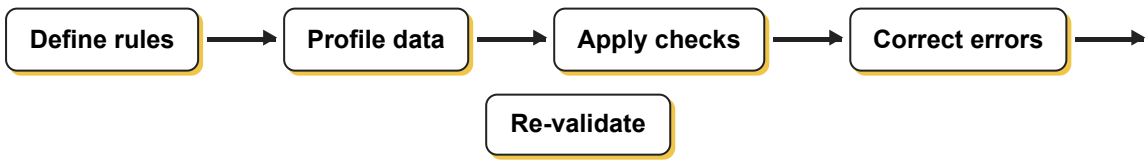
Q1. What tools are used for data validation? What are the important steps of data validation? (Jan 2023)

Data validation tools:

- **Spreadsheet validation:** Excel/Google Sheets data validation rules, drop-down lists, conditional formatting.
- **Database constraints:** NOT NULL, UNIQUE, PRIMARY/FOREIGN KEY, CHECK, triggers and SQL queries.
- **ETL tools:** Informatica, Talend, SSIS, Pentaho with built-in validation transformations.
- **Data quality and profiling tools:** OpenRefine, Trifacta, Talend Data Quality, Great Expectations.
- **Programming libraries:** Python (pandas, Cerberus, pydantic), R packages.
- **BI tools:** Power BI and Tableau data profiling features.

Important steps:

1. **Define rules and requirements:** data types, ranges, formats and business rules.
2. **Data profiling:** examine data to understand structure, nulls and patterns.
3. **Apply checks:** data type, range, format, length, uniqueness, presence (missing values), consistency (cross-field) and referential integrity.
4. **Identify and log errors:** flag invalid records.
5. **Correct or reject:** clean, standardise, fill or remove bad data.
6. **Re-validate and document:** confirm fixes and record results.
7. **Automate and monitor** continuously.



Data validation steps

SECTION C Answers (≈500 Words Each)

Q1. What is the function to find the day of the week for a particular date value? (Jan 2023)

Most tools provide a function:

- **Excel / Google Sheets:** `=WEEKDAY(A1)` returns 1 (Sunday) to 7 (Saturday); `=WEEKDAY(A1, 2)` returns 1 (Monday) to 7 (Sunday). `=TEXT(A1, "dddd")` returns the name, e.g. Monday.
- **SQL (MySQL):** `DAYOFWEEK(date)` or `DAYNAME(date)`; SQL Server: `DATENAME(weekday, date)`; Oracle: `TO_CHAR(date, 'DAY')`.
- **Python (pandas):** `df['d'].dt.day_name()` or `date.weekday()` (Monday = 0).
- **Power BI (DAX):** `WEEKDAY([Date])` or `FORMAT([Date], "dddd")`.

Example: `=WEEKDAY("2023-01-15")` returns 1, meaning Sunday.

Q2. What is data validation? How does it work and why is it important? (Jan 2023)

Data validation is the process of checking data for **accuracy, completeness, consistency and format** before it is used, stored or analysed.

How it works:

1. Validation rules are defined (type, range, format, uniqueness, required fields, business rules).
2. Data is profiled to understand its current state.
3. Each record is checked against the rules at entry (front-end forms, database constraints) or during ETL.
4. Invalid records are flagged, logged, and either corrected, standardised or rejected.
5. The data is re-checked and validation is repeated regularly.

Types of checks:

Check	Example
Type	Age must be numeric
Range	Age between 0 and 120
Format	Email contains @, date as DD-MM-YYYY
Uniqueness	No duplicate customer ID
Presence	Mobile number not blank
Referential	Order's customer ID exists in customer table
Consistency	Delivery date after order date



How data validation works

Importance:

- Ensures reliable analytics and reports; bad input gives bad decisions ("garbage in, garbage out").
- Saves time and cost of fixing errors later.
- Maintains data integrity and consistency across systems.
- Helps meet compliance and regulatory requirements.
- Builds user trust in BI results and improves customer experience.

Conclusion: data validation is the first quality gate of every BI and analytics project.
