Shortly :

Redoing data collection in data preparation is necessary to address issues like poor data quality (incomplete, inaccurate, or irrelevant data), changing project requirements, sampling bias, technical errors, regulatory compliance, or the need for more scalable or representative data. It ensures the dataset is reliable, complete, and aligned with the analysis goals, leading to more accurate and meaningful results.

---

Redoing data collection during the data preparation phase is often necessary for several reasons:

**1. Poor Data Quality**

- **Incomplete Data**: If the initial dataset has missing values or gaps, it may not be sufficient for analysis or modeling.
- **Inaccurate Data**: Errors, inconsistencies, or outdated information in the dataset can lead to incorrect conclusions.
- **Irrelevant Data**: The data collected might not align with the problem being solved or the objectives of the analysis.

**2. Changing Requirements**

- **Evolving Goals**: The objectives of the project may change over time, requiring additional or different data.
- **New Insights**: Preliminary analysis might reveal that the initial dataset lacks critical variables or features needed for deeper insights.

**3. Bias or Unrepresentative Data**

- **Sampling Bias**: The initial dataset might not represent the population or phenomenon being studied, leading to biased results.
- **Underrepresented Groups**: Certain groups or scenarios might be missing, making the dataset unsuitable for generalization.

**4. Technical Issues**

- **Data Collection Errors**: Mistakes in the data collection process (e.g., faulty sensors, incorrect survey questions) can compromise the dataset.
- **Storage or Transfer Issues**: Data might be corrupted or lost during storage or transfer, necessitating re-collection.

**5. Regulatory or Ethical Concerns**

- **Compliance Issues**: The initial data collection might not comply with legal or ethical standards (e.g., GDPR, privacy laws), requiring a redo.
- **Consent Issues**: If proper consent was not obtained for data usage, re-collection might be necessary.

**6. Improved Methods or Tools**

- **Better Technology**: New tools or methods for data collection might become available, offering higher accuracy or efficiency.
- **Enhanced Processes**: Lessons learned from the initial collection process can lead to improved methodologies for re-collection.

**7. Scalability Needs**

- **Insufficient Volume**: The initial dataset might be too small to support robust analysis or machine learning models.
- **Expanded Scope**: The project might expand to include additional variables, time periods, or geographic regions.

**8. Validation and Verification**

- **Cross-Validation**: Re-collecting data can help validate the findings from the initial dataset, ensuring reliability.
- **Reproducibility**: Repeating data collection ensures that results are reproducible and not due to random chance or anomalies in the initial dataset.

**9. Feedback from Stakeholders**

- **Stakeholder Input**: Feedback from stakeholders might highlight gaps or issues in the initial dataset, prompting a redo.
- **Alignment with Business Needs**: The data might need to be re-collected to better align with business objectives or decision-making processes.

**10. Time Sensitivity**

- **Outdated Data**: If the initial data is time-sensitive (e.g., market trends, seasonal data), it might need to be re-collected to reflect current conditions.

**Conclusion**

Redoing data collection is a critical step to ensure the dataset is accurate, complete, relevant, and aligned with the project's goals. While it can be time-consuming and costly, it is often necessary to achieve reliable and meaningful results. Proper planning and evaluation during the initial data collection phase can minimize the need for redoing it, but flexibility and adaptability are key in data-driven projects.