IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 1, November 2026
Impact Factor: 8.2

Data Quality and Incremental Data Loading in ETL Pipelines: A Comparative Survey of Industry Practices and Student Readiness

Sanket Sanjay Jadhav, Rohan Murlidhar Mule, Rupali Lahu Pawar,
Prof. Jayshri Thorat, Dr. Ranmalkar V.S
Department of Master of Computer Applications (MCA)
Adsul's Technical Campus, Chas, Ahilyanagar, Maharashtra, India

Abstract : *Extract–Transform–Load (ETL) pipelines depend on data quality validation, reconciliation and correct identification of changed records. This cross-sectional questionnaire survey compares practices reported by working professionals with the awareness and practical exposure of IT students. Of 90 submitted responses, 84 were analysed (34 professionals, 50 students) after excluding non-consenting and unclassifiable responses. Professionals most often reported null checks, format validation and duplicate detection, and incremental loading was the most common primary approach. Students reported lower familiarity with incremental loading, data reconciliation and Change Data Capture (CDC) and limited project-based exposure. Professionals reported higher ETL familiarity than students (Mann–Whitney $U = 1179$, $p = 0.00164$). The study is exploratory and relies on self-reported data from a convenience sample, so its findings should not be generalised.*

Keywords: ETL, data quality, incremental loading, change data capture, data engineering education, survey

I. INTRODUCTION

Organisations use ETL pipelines to extract data from operational sources, transform it into consistent structures and load it into analytical repositories. The reliability of these pipelines depends not only on successful data movement but also on validation, duplicate handling, schema consistency, reconciliation and the correct identification of changed records, because incomplete or inconsistent data can affect downstream reports and decisions. ETL design and data-quality management are long-standing concerns in data warehousing [1], [2]. Fig. 1 summarises a generic ETL workflow.

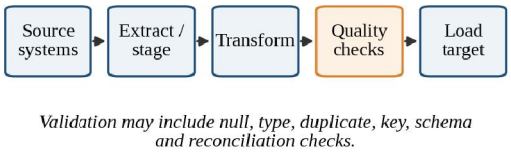


Fig. 1. Conceptual ETL workflow and representative quality checks. Source: author-created illustration informed by [1], [2], [3], [4].

Two recurring engineering concerns are data quality management and the choice between full reload and incremental loading. A full reload processes an entire dataset, whereas incremental loading processes only new or changed records. Incremental approaches can reduce repeated processing, but they depend on reliable change identification and on careful handling of updates, deletes and late-arriving records. Post et al. examined the trade-offs between full reload and incremental updating in a research data warehouse [5].

Copyright to IJARSCT
www.ijarsct.co.in
DOI: 10.48175/IJARSCT-40550

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 1, November 2026
Impact Factor: 8.2

A second concern is the gap between academic exposure and practical data engineering work. Students may study SQL, Python and data cleaning without gaining equivalent experience in reconciliation, change data capture (CDC) or the maintenance of multi-stage pipelines. This study therefore surveys both working professionals and IT students.

A. Research Problem

This study aims to understand which data quality and loading practices are reported by professionals and how familiar students are with these concepts. A comparative survey can indicate areas where practical data engineering education may need greater emphasis.

B. Objectives

The objectives are (i) to examine data quality practices reported by working data professionals; (ii) to identify commonly reported validation and reconciliation checks; (iii) to study reported incremental loading and source-change identification approaches; (iv) to identify challenges encountered in ETL pipeline maintenance; (v) to assess students' awareness and practical exposure to ETL and data quality concepts; and (vi) to compare self-reported ETL familiarity between students and professionals.

C. Research Questions

The research questions are: (RQ1) Which data quality checks are reported by working professionals? (RQ2) Which loading and change-detection approaches are used in their projects? (RQ3) What ETL maintenance challenges are reported? (RQ4) How familiar are students with data quality, reconciliation and incremental loading? (RQ5) How does self-reported ETL familiarity differ between students and professionals?

II. METHODOLOGY

A. Research Design

The study follows a cross-sectional, descriptive and comparative survey design. Primary data were collected through a structured Google Forms questionnaire containing common questions for all respondents and separate sections for working professionals and students. The screening and analysis sequence is summarised in Fig. 2.

B. Population and Sampling

The intended respondent groups were (i) IT students enrolled in undergraduate or postgraduate programmes and (ii) working professionals in data-related or technology roles. Convenience sampling was used. The form was circulated through college networks, WhatsApp groups and personal contacts. Participation was voluntary.

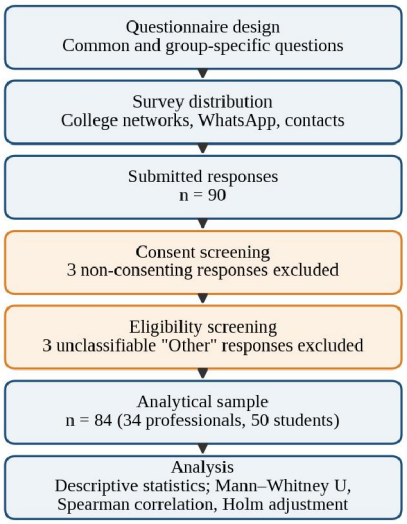


Fig. 2. Survey screening and analysis workflow (n=90 submitted; n=84 included)

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 1, November 2026
Impact Factor: 8.2

C. Sample Selection

Table I and Fig. 2 summarise the response disposition, and Table II gives the respondent profile. Only responses with consent and a recognised student or working-professional category were retained for the main analysis. Question-level denominators were used where optional or branch-specific responses were missing.

TABLE I: SAMPLE SELECTION

Response disposition	Number
Total submitted responses	90
Did not consent	3
Consented; category not classifiable into the two study groups	3
Eligible analytical sample	84
Working professionals	34
Students (undergraduate and postgraduate)	50

TABLE II: RESPONDENT PROFILE

Respondent category	N	Percentage
Working professionals	34	40.5
Postgraduate students	30	35.7
Undergraduate students	20	23.8
Total	84	100.0

D. Questionnaire and Variables

The questionnaire collected respondent category, area of study or work, familiarity with ETL, technologies used or studied, and familiarity with selected data engineering concepts. The professional section covered job role, experience, tools, data quality checks, automation, loading strategy, change detection, pipeline challenges and reconciliation. The student section covered academic level, curriculum exposure, practical data cleaning experience, technology exposure, familiarity with Medallion Architecture, confidence and perceived barriers to practical learning. The full questionnaire is provided as supplementary material.

E. Data Analysis

Responses were analysed using Python with Pandas and SciPy. Frequencies and percentages were calculated for categorical and multiple-selection questions (each option counted independently, so percentages may exceed 100%; no missing values were imputed), and five-point familiarity responses were coded from 1 (Not familiar) to 5 (Extremely familiar). The Mann–Whitney U test compared ETL familiarity scores between students and professionals, with the rank-biserial correlation as an effect size. A second Mann–Whitney U test compared professional experience ranks between two grouped loading-approach categories. Spearman's rank correlation examined the association between student ETL familiarity and practical exposure to data cleaning and validation. For the eight common concept-familiarity items (Q6), concept-wise Mann–Whitney tests were adjusted for multiple comparisons using the Holm method, and an exploratory Mann–Whitney comparison was made between undergraduate and postgraduate students. All tests were two-sided with a 5% significance level.

F. Research Hypotheses

H01: There is no significant difference in self-reported ETL familiarity between students and working professionals.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 1, November 2026 Impact Factor: 8.2

H02: There is no significant difference in professional experience ranks between respondents reporting incremental loading/CDC and respondents reporting full reload or combined use.
H03: There is no significant association between students' self-reported ETL familiarity and their ordered practical exposure to data cleaning and validation.

G. Ethical Considerations

Participation was voluntary, and respondents were asked to provide informed consent before proceeding. The three responses indicating that consent was not given were excluded. The questionnaire was administered for academic research, and responses are reported in aggregate. No personally identifying respondent information is presented in this paper.

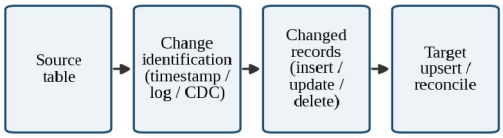
III. LITERATURE REVIEW

Data loading and ETL design are long-standing data warehouse concerns. Vassiliadis surveys ETL technology, covering the conceptual and logical modelling of ETL processes, the extraction, transformation and loading stages, and problems that affect the process as a whole [1]. Kimball and Caserta describe practical ETL design tasks, including extraction from multiple sources, cleaning, conforming and delivering data to analytical structures [2]. Post et al. described the implementation of daily incremental updating for an i2b2 research data warehouse and discussed its trade-offs against periodic full reloads, including the burden on source systems, the up-front development effort and the fit with organisational structures [5]. Zahid et al. conducted a systematic literature review of big data analytics in telecommunications and proposed an architecture, LambdaTel, for practitioners [6]; this provides sector context but does not address the practices examined in the present survey.

Data quality research treats quality as broader than accuracy. Wang and Strong derived a hierarchical framework from data consumers' perspectives with four categories: intrinsic, contextual, representational and accessibility data quality [7]. Batini et al. compare data quality assessment and improvement methodologies along dimensions that include their phases and steps, strategies and techniques, and data quality dimensions [3]. Rahm and Do classify the data quality problems addressed by data cleaning, distinguishing single-source from multi-source problems, and note that cleaning is especially needed when heterogeneous sources are integrated [4]. For ETL validation, Homayouni et al. proposed automated balancing tests and defined three categories of properties to be checked: completeness, consistency and syntactic validity [8]. These categories are conceptually related to the survey items on record-count reconciliation, null checks, duplicate detection, schema validation and format validation, and the studies above provide background for the data quality items in the questionnaire.

Incremental processing requires the system to identify records that are new or changed instead of reloading all data. Jörg and Deßloch examine the generation of ETL processes for incremental loading [9]. In data lake environments, Armbrust et al. describe the Lakehouse architecture, which aims to combine the low-cost, open storage of data lakes with data-warehouse management features [10], and the Delta Lake paper addresses ACID transactions for mutable tables on cloud object stores [11]. These publications provide technical background for the survey's questions on change detection, incremental loading and updates; they do not imply that respondents use a lakehouse or a particular CDC implementation. Fig. 3 illustrates the generic logic of incremental loading, and Fig. 4 shows the layered Bronze–Silver–Gold pattern referred to in the survey's Medallion-architecture question.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 1, November 2026
Impact Factor: 8.2



Conceptual illustration; implementations differ by source and platform.

Fig. 3. Conceptual incremental loading and change-data processing workflow. Source: author-created illustration informed by [9], [11].



Conceptual layered-data pattern; layer definitions vary across implementations.

Fig. 4. Conceptual Bronze–Silver–Gold data-layer pattern. Source: author-created illustration

The study uses self-reported five-point familiarity ratings. Sullivan and Artino note that Likert-type responses are ordinal and discuss the long-standing debate over whether such data can be treated as interval data in parametric analysis [12]. Because the ratings are ordinal and the groups are small, the present study uses rank-based non-parametric tests, which that article lists among the approaches for ordinal data. Etikan et al. describe convenience sampling as a non-probability technique and discuss its limitations for representing a population [13]. The Mann–Whitney U statistic was proposed by Mann and Whitney [14], who note that an equivalent test had earlier been proposed by Wilcoxon.

Although the cited literature covers ETL design, data quality, incremental processing and survey methods, the present study addresses a narrower empirical question: which practices does this convenience sample of professionals report, and how do students describe their exposure and familiarity? The survey is descriptive and exploratory and is not designed to estimate industry-wide adoption or establish causal effects.

IV. RESULTS AND DISCUSSION

A. Respondent Profile

The analytical sample consisted of 84 respondents: 34 working professionals (40.5%), 30 postgraduate students (35.7%) and 20 undergraduate students (23.8%). Students therefore represented 59.5% of the retained sample (Table II).

B. Reported Professional Practices

Among the 34 professionals, null or missing-value checks were selected most often (22; 64.7%), followed by data type and format validation (20; 58.8%) and duplicate detection (18; 52.9%) (Fig. 5). Referential-integrity checks were selected by six respondents (17.6%), and two respondents (5.9%) reported no regular data quality checks. Multiple selections were permitted.

Regarding automation, 13 professionals (38.2%) reported checks that were mostly automated with manual intervention, six (17.6%) fully automated checks, 11 (32.4%) mostly manual checks and two (5.9%) completely manual checks; two respondents (5.9%) were unsure (Fig. 6).

C. Loading and Change Detection

Incremental loading was selected as the primary approach by 15 professionals (44.1%). Eight (23.5%) selected full reload, six (17.6%) reported using both approaches depending on requirements, two (5.9%) selected CDC as the primary approach and three (8.8%) were unsure (Fig. 7).

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 1, November 2026 Impact Factor: 8.2

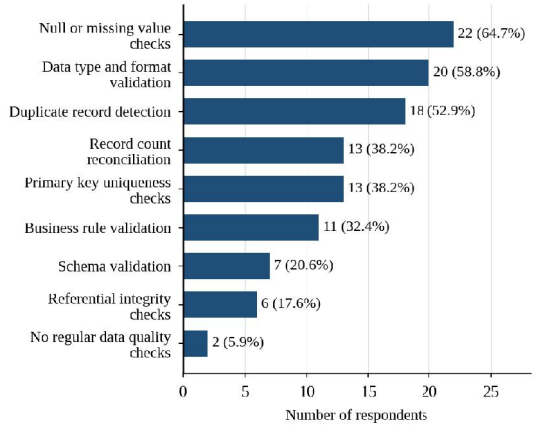


Fig. 5. Data quality checks reported by working professionals (n=34; multiple responses allowed). Labels show count and percentage of professionals.

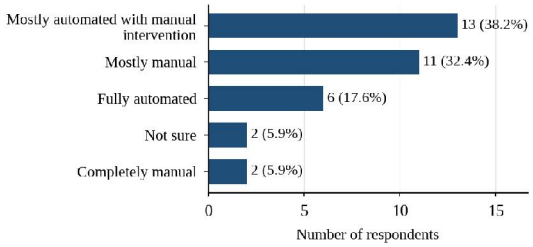


Fig. 6. Reported implementation of data quality checks (professionals; n=34).

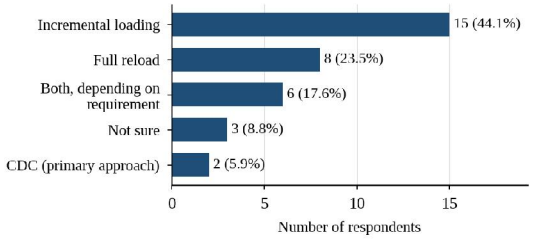


Fig. 7. Primary loading approach reported by working professionals (n=34).

Timestamp or last-updated columns and complete-dataset comparison were the most frequently selected change-identification methods (eight respondents each; 23.5%), followed by database CDC (six; 17.6%) (Fig. 8). Three respondents were unsure and one selected "Other". These responses describe reported methods and should not be interpreted as a measure of technical maturity.

D. Reported Pipeline Challenges

Fig. 9 shows the challenges reported by professionals. Missing or null values and inconsistent data formats were each selected by 18 professionals (52.9%), followed by duplicate records (15; 44.1%) and failed pipeline executions (14; 41.2%). These results indicate that data correctness and operational reliability were relevant concerns among the surveyed professionals.

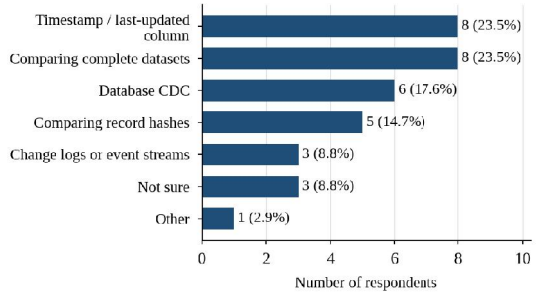


Fig. 8. Reported source-change identification methods (professionals; n=34).

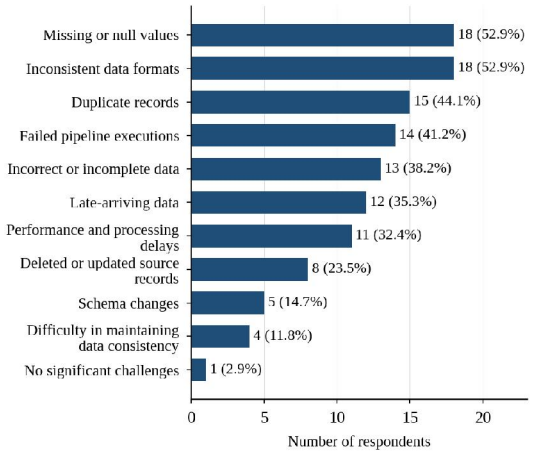


Fig. 9. ETL pipeline challenges reported by working professionals (n=34; multiple responses allowed).

Copyright to IJARSCT
www.ijarsct.co.in
DOI: 10.48175/IJARSCT-40550

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 1, November 2026
Impact Factor: 8.2

E. Student Awareness and Exposure

Among 50 students, 20 (40.0%) reported a basic curricular introduction to ETL, 17 (34.0%) had learned about it only through online courses or self-learning, nine (18.0%) reported no curricular exposure and four (8.0%) had studied ETL in detail. For practical data cleaning and validation, 17 (34.0%) reported classroom exercises only, 12 (24.0%) theoretical knowledge only, 10 (20.0%) one project, seven (14.0%) no exposure and four (8.0%) multiple projects (Fig. 10). Regarding the Medallion architecture (Fig. 11), 19 students (38.0%) had never heard of it, 13 (26.0%) had heard the terms without understanding them, nine (18.0%) had limited understanding and nine (18.0%) understood how it worked.

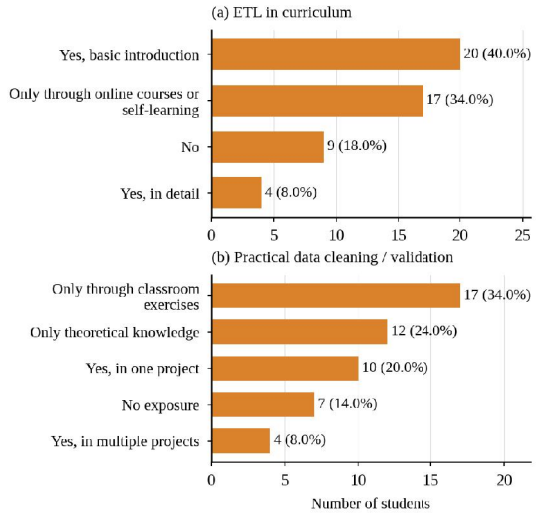


Fig. 10. Student-reported ETL curriculum exposure (a) and practical data cleaning or validation exposure (b) (n=50).

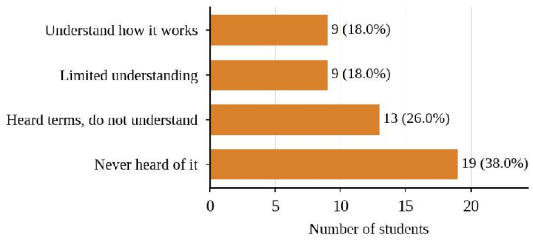


Fig. 11. Student awareness of Medallion architecture (n=50).

A further exploratory comparison used a question common to all respondents (Q6 in the supplementary questionnaire), which asked all 84 respondents to rate their familiarity with the same eight ETL concepts (Fig. 12). Working professionals reported higher mean familiarity for all eight concepts. The largest mean differences were observed for CDC (2.24 for professionals and 1.54 for students) and data reconciliation (2.74 and 2.10, respectively).

Unadjusted Mann–Whitney tests showed group differences for seven concepts ($p < 0.03$); duplicate detection was not significant ($p = 0.120$). Because eight concept-wise tests were conducted, Holm adjustment was also applied: CDC (adjusted $p = 0.018$) and data reconciliation (adjusted $p = 0.034$) remained below 0.05, whereas the remaining differences did not. These comparisons are exploratory and should not be interpreted as eight independent confirmatory findings.

Students also rated eight student-specific items. Fig. 13(a) shows that students reported the lowest familiarity with CDC (mean 1.66), incremental loading (1.92) and data reconciliation (1.98). The most frequently selected barriers to practical experience (Fig. 13(b)) were lack of practical projects (29; 58.0%), lack of proper guidance (21; 42.0%), limited access to cloud platforms (19; 38.0%) and difficulty understanding ETL concepts (18; 36.0%). Multiple selections were allowed.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 1, November 2026
Impact Factor: 8.2

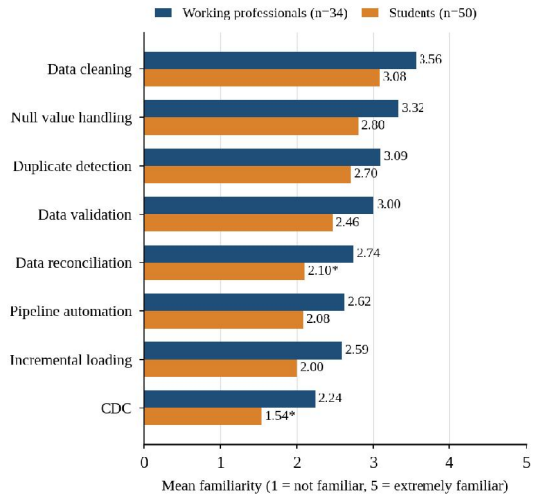


Fig. 12. Mean familiarity with eight common ETL concepts (professionals n=34; students n=50). * Significant after Holm adjustment (adjusted $p < 0.05$).

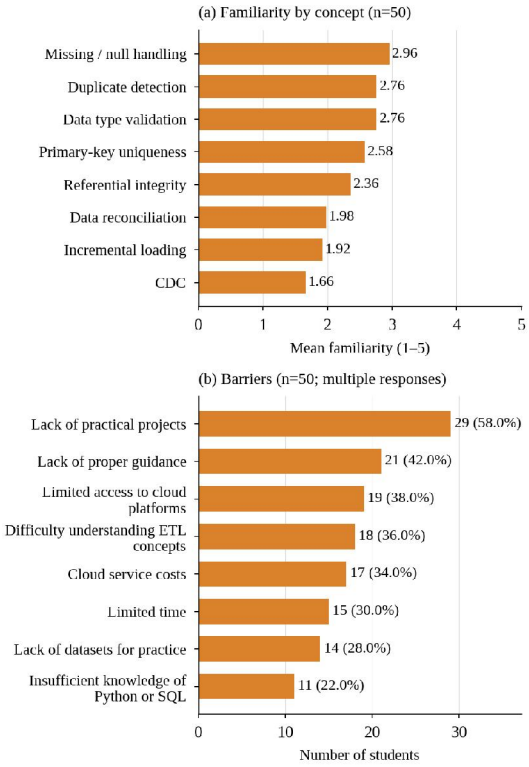


Fig. 13. Student familiarity by ETL concept (a) and reported barriers to practical experience (b) (n=50).

Fig. 14 shows the distribution of both ratings. The mean confidence in independently building a basic ETL pipeline was 2.22 out of 5, whereas 43 of 50 students (86.0%) agreed or strongly agreed that hands-on projects would improve their data engineering readiness. This contrast suggests that respondents perceive a need for practical learning opportunities; it does not demonstrate that a particular intervention would improve skills.

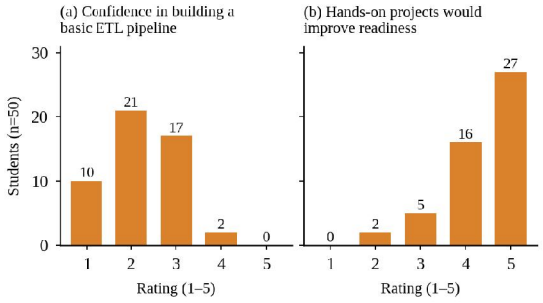


Fig. 14. Student confidence in building a basic ETL pipeline (a) and perceived value of hands-on projects (b) (n=50).

F. Hypothesis Testing

1) H01: Difference in ETL Familiarity

The Mann–Whitney U test returned $U = 1179$ and $p = 0.00164$. Professionals had a mean score of 3.18 ($n = 34$; median = 3), while students had a mean score of 2.44 ($n = 50$; median = 2). At the 5% significance level, H01 is rejected. The rank-biserial effect size was approximately 0.39 in favour of professionals.

Copyright to IJARSCT
www.ijarsct.co.in
DOI: 10.48175/IJARSCT-40550

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 1, November 2026 Impact Factor: 8.2

2) H02: Experience and Loading Approach

For an exploratory comparison, incremental loading and CDC were grouped together (n = 17), while full reload and combined use were grouped together (n = 14). The three "Not sure" responses were excluded. The Mann–Whitney U test returned U = 90 and p = 0.2408. The mean coded experience score was 2.29 for the incremental/CDC group and 2.71 for the full-reload/combined group. The result is not statistically significant, so H02 is not rejected. The direction is noteworthy: the incremental/CDC group had lower mean experience in this sample, contrary to a possible expectation that more experienced respondents would be more likely to use incremental approaches. The subgroup counts are small, so this pattern is indicative only.

3) H03: Student Familiarity and Practical Exposure

Spearman's rank correlation used all 50 student responses, including the seven respondents who selected "No exposure". Practical exposure was coded ordinally as 1 = no exposure, 2 = theoretical knowledge only, 3 = classroom exercises only, 4 = one project and 5 = multiple projects. The correlation was $\rho = 0.0336$, $p = 0.8169$. The association was not statistically significant; H03 is not rejected. The ordering of theoretical knowledge and classroom exercises is a judgement-based coding decision, and alternative coding could produce a different estimate. The result does not establish that familiarity and practical exposure are unrelated in the wider student population.

4) Hypothesis Summary

Table III summarises the three hypothesis tests.

TABLE III: SUMMARY OF HYPOTHESIS TESTS

Hypothesis	Test result	Decision
H01: No difference in ETL familiarity between groups	U = 1179; p = 0.00164	Reject H01
H02: No difference in experience ranks by loading approach	U = 90; p = 0.2408	Do not reject H02
H03: No association between student familiarity and practical exposure	ρ = 0.0336; p = 0.8169	Do not reject H03

G. Discussion

The professional responses indicate that routine validation practices such as null checks, format validation and duplicate detection were reported more often than referential-integrity checks. This pattern suggests that foundational checks are prominent in the surveyed projects, whereas some relational and reconciliation checks may receive less consistent attention; the survey does not establish why. The distributions are shown in Figs. 5–9.

Incremental loading was the most frequently selected primary approach, although full reload and combined strategies were also reported. The presence of multiple approaches is consistent with the idea that loading strategy depends on dataset characteristics and operational needs. The survey did not collect enough detail to assess whether a particular strategy performed better.

Student responses show a contrast between exposure to foundational tools and familiarity with advanced pipeline concepts. SQL and Python were commonly selected technologies, whereas familiarity with CDC, incremental loading and reconciliation was lower than for foundational tasks such as data cleaning and null handling, and mean confidence in building a basic ETL pipeline was modest. A basic curricular introduction was the most common formal exposure and classroom exercises the most common practical experience, which suggests greater exposure to introductory data preparation than to operational pipeline maintenance among these respondents. The high level of agreement that hands-on projects would improve readiness suggests that respondents perceive practical learning opportunities as valuable. These interpretations apply to the respondents only and should not be generalised to all MCA or IT programmes (Figs. 10–14).

The significant difference in self-rated ETL familiarity between professionals and students should be interpreted cautiously. The groups may apply different internal standards when selecting a point on the 1–5 scale, so the measure is not necessarily equivalent across groups. The cross-sectional design also cannot establish that professional experience caused the difference.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 1, November 2026 Impact Factor: 8.2

An additional exploratory comparison between undergraduate (n = 20; mean ETL familiarity = 2.25) and postgraduate students (n = 30; mean = 2.57) was not statistically significant (Mann–Whitney U = 241, p = 0.2143). This comparison is also limited by the modest subgroup sizes.

Self-reported project experience also requires caution. Eleven of 50 students selected "Yes, implemented a complete pipeline," compared with four of 34 professionals. This unexpected pattern should not be interpreted as evidence that students have more extensive pipeline experience: respondents may interpret "complete pipeline" differently, and the survey did not verify project scope or implementation responsibility.

H. Implications

For education, practical coursework could include end-to-end exercises covering source extraction, transformation, validation, duplicate handling, reconciliation, incremental updates and failure recovery. Students could compare full reload and incremental loading using controlled datasets and document the trade-offs.

For practice, the responses reinforce the relevance of repeatable validation checks and reconciliation. Teams may benefit from defining explicit quality rules, recording validation outcomes and documenting how changed or late-arriving records are handled. These are implications suggested by the survey themes, not experimentally validated recommendations.

I. Limitations and Future Work

The study has several limitations. First, convenience sampling through college networks, WhatsApp groups and personal contacts may introduce selection bias, and the sample is not statistically representative of all IT students or data professionals. Second, the professional sample contains 34 respondents, and experience subgroups are small (eight respondents had less than one year of experience and five had more than five years), so subgroup patterns are indicative only. Third, all findings are self-reported and may be affected by differing interpretations of terms such as ETL, automation and CDC; self-rated familiarity may also not be directly comparable between students and professionals because the groups may use the 1–5 scale differently. Fourth, H02 combines incremental loading and CDC into one category for a small-sample comparison, although they are distinct methods. Fifth, the eight concept-wise comparisons in Fig. 12 are exploratory and subject to multiple-testing concerns, so Holm-adjusted values are reported. Finally, the questionnaire did not directly measure the frequency of data quality incidents, so a hypothesis relating automation to incident frequency could not be tested.

Future research could recruit a larger and more diverse professional sample, refine the questionnaire through cognitive interviews, measure incident frequency directly, distinguish incremental loading from CDC more precisely, and collect evidence from practical ETL exercises or interviews. A broader sampling frame would improve the generalisability of findings.

V. CONCLUSION

This survey examined reported ETL data quality and loading practices among 34 working professionals and the awareness and readiness of 50 IT students. Professionals most frequently reported null checks, data-format validation and duplicate detection. Incremental loading was the most frequently selected primary loading approach. Students reported lower familiarity with CDC, incremental loading and data reconciliation, as well as modest confidence in independently building ETL pipelines. The comparison found a statistically significant difference in ETL familiarity between professionals and students, while the two exploratory within-group tests were not statistically significant. Given the convenience sample and self-reported nature of the data, the findings should be treated as exploratory. The study highlights practical data engineering exercises as a potential area for strengthening student readiness.

ACKNOWLEDGMENTS

The authors thank the students and working professionals who took part in the survey.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 1, November 2026
Impact Factor: 8.2

REFERENCES

[1] P. Vassiliadis, "A survey of extract–transform–load technology," Int. J. Data Warehousing Mining, vol. 5, no. 3, pp. 1–27, Jul.–Sep. 2009, doi: 10.4018/jdwm.2009070101.
[2] R. Kimball and J. Caserta, The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data. Indianapolis, IN, USA: Wiley, 2004.
[3] C. Batini, C. Cappiello, C. Francalanci, and A. Maurino, "Methodologies for data quality assessment and improvement," ACM Comput. Surveys, vol. 41, no. 3, Art. no. 16, pp. 1–52, Jul. 2009, doi: 10.1145/1541880.1541883.
[4] E. Rahm and H. H. Do, "Data cleaning: Problems and current approaches," IEEE Data Eng. Bull., vol. 23, no. 4, pp. 3–13, Dec. 2000.
[5] A. R. Post, M. Ai, A. K. Pai, M. Overcash, and D. S. Stephens, "Architecting the data loading process for an i2b2 research data warehouse: Full reload versus incremental updating," AMIA Annu. Symp. Proc., vol. 2017, pp. 1411–1420, 2018.
[6] H. Zahid, T. Mahmood, A. Morshed, and T. Sellis, "Big data analytics in telecommunications: Literature review and architecture recommendations," IEEE/CAA J. Autom. Sinica, vol. 7, no. 1, pp. 18–38, Jan. 2020, doi: 10.1109/JAS.2019.1911795.
[7] R. Y. Wang and D. M. Strong, "Beyond accuracy: What data quality means to data consumers," J. Manage. Inf. Syst., vol. 12, no. 4, pp. 5–33, 1996, doi: 10.1080/07421222.1996.11518099.
[8] H. Homayouni, S. Ghosh, and I. Ray, "An approach for testing the extract-transform-load process in data warehouse systems," in Proc. 22nd Int. Database Eng. Appl. Symp. (IDEAS), 2018, pp. 236–245, doi: 10.1145/3216122.3216149.
[9] T. Jörg and S. Deßloch, "Towards generating ETL processes for incremental loading," in Proc. 12th Int. Database Eng. Appl. Symp. (IDEAS), Coimbra, Portugal, Sep. 2008, pp. 101–110.
[10] M. Armbrust, A. Ghodsi, R. Xin, and M. Zaharia, "Lakehouse: A new generation of open platforms that unify data warehousing and advanced analytics," in Proc. 11th Annu. Conf. Innov. Data Syst. Res. (CIDR), Online, Jan. 2021.
[11] M. Armbrust et al., "Delta Lake: High-performance ACID table storage over cloud object stores," Proc. VLDB Endowment, vol. 13, no. 12, pp. 3411–3424, 2020, doi: 10.14778/3415478.3415560.
[12] G. M. Sullivan and A. R. Artino Jr., "Analyzing and interpreting data from Likert-type scales," J. Graduate Med. Educ., vol. 5, no. 4, pp. 541–542, Dec. 2013, doi: 10.4300/JGME-5-4-18.
[13] I. Etikan, S. A. Musa, and R. S. Alkassim, "Comparison of convenience sampling and purposive sampling," Amer. J. Theor. Appl. Statist., vol. 5, no. 1, pp. 1–4, 2016, doi: 10.11648/j.ajtas.20160501.11.
[14] H. B. Mann and D. R. Whitney, "On a test of whether one of two random variables is stochastically larger than the other," Ann. Math. Statist., vol. 18, no. 1, pp. 50–60, Mar. 1947, doi: 10.1214/aoms/1177730491.