

An Overview of Data Science and ETL Operations: Internship Experience at Cliqpack Limited

by

Monira Akter (221014039)

*Internship report (CSE4099) submitted in partial fulfillment of the
requirements for the degree of*

Bachelor of Science in Computer Science and Engineering

Under the supervision of

Raihan Kibria

Senior Lecturer, Department of Computer Science and Engineering, University of
Liberal Arts Bangladesh

Md.Ahsanul Haq

Senior Data Engineer, Cliqpack Limited



**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
UNIVERSITY OF LIBERAL ARTS BANGLADESH**

SPRING 2026

© *Monira Akter*
All rights reserved

DECLARATION

Internship Title	An Overview of Data Science and ETL Operations: Internship Experience at Cliqpack Limited
Author	<i>Monira Akter</i>
Student ID	221014039
Supervisor (Academic)	Raihan Kibria
Supervisor (Industry)	Md.Ahsanul Haq

I declare that this internship report entitled *An Overview of Data Science and ETL Operations: Internship Experience at Cliqpack Limited* is the result of my own work except as cited in the references. The internship report has not been accepted for any degree and is not concurrently submitted in candidature of any other degree.

Monira Akter
221014039

Department of Computer Science and Engineering
University of Liberal Arts Bangladesh

Date: April 21, 2026



Department of Computer Science and Engineering
University of Liberal Arts Bangladesh
Mohammadpur, Dhaka - 1207

CERTIFICATE

This is to certify that the internship report entitled **An Overview of Data Science and ETL Operations: Internship Experience at Cliqpack Limited**, submitted by **Monira Akter** (Student ID: 221014039) is a undergraduate student of the **Department of Computer Science and Engineering** has been examined. Upon recommendation by the examination committee, we hereby accord our approval of it as the presented work and submitted report fulfill the requirements for its acceptance in partial fulfillment for the degree of *Bachelor of Science in Computer Science and Engineering*.

**Raihan Kibria, Senior Lecturer, Department of
Computer Science and Engineering, University of
Liberal Arts Bangladesh**

**Md.Ahsanul Haq , Senior Data Engineer, Cliqpack
Limited**

Place: Dhaka

Date: April 21, 2026

ACKNOWLEDGEMENTS

I would like to express my deep and sincere gratitude to my industry supervisor, *Md.Ahsanul Haq*, and my academic supervisor *Raihan Kibria* for allowing me to conduct this internship and providing invaluable guidance throughout this work. Their dynamism, vision, sincerity, and motivation have deeply inspired me. They have taught me the methodology to carry out the work and to present the work as clearly as possible. It was a great privilege and honor to work and study under his guidance. Also, I would like to thank the other employees of *Cliqpack Limited* for their help and cooperation during my practice time.

I am greatly indebted to our honorable teachers of the Department of Computer Science and Engineering at the University of Liberal Arts Bangladesh who taught me during the course of our study. Without any doubt, their teaching and guidance have completely transformed us to the persons that we are today.

I am extremely thankful to my parents for their unconditional love, endless prayers and caring, and immense sacrifices for educating and preparing me for the future. I would like to say thanks to my friends and relatives for their kind support and care.

Finally, I would like to thank all the people who have supported me to complete the Internship directly or indirectly.

Monira Akter

University of Liberal Arts Bangladesh

Date: April 21, 2026

Dedicated to
my loving Parents

– *Monira Akter*

EXECUTIVE SUMMARY

The internship at Cliqpack Limited provided me with an opportunity to come out of classroom theory into real world, practical network operations. Cliqpack Limited is one of the most established IT companies in Bangladesh, which was founded in 2017 in Australia. The organization offers a wide range of data-driven and technology based services, including data collection and processing solutions, web scraping and data extraction services, data pipeline development (ETL), data validation and preprocessing, dataset management, research and development support, and other data science and analytics-related services for corporate and business clients.

As an intern in the Data Science Engineering Department, my responsibility was to learn about data-driven processes and understand how data is collected, processed, and managed within the organization. These activities included extracting data using web scraping techniques such as XPath and CSS selectors, building and managing data pipelines (ETL), performing data validation and preprocessing, cleaning and merging datasets by handling duplicates and missing values, and conducting research and development by sourcing public datasets and planning project timelines. My role primarily focused on learning and supporting these data engineering operations under supervision, rather than independently managing them.

A significant component of the internship involved accompanying senior engineers. These experiences enabled me to observe real-world technical problems, understand their root causes, and learn appropriate professional conduct and communication etiquette when interacting with clients in a corporate environment.

Overall, the internship provided valuable insight into data science engineering processes and data management operations while contributing to the development of my technical knowledge, analytical skills, problem-solving abilities, and professional competence.

Keywords: Data Science Engineering, Web Scraping, Data Extraction, ETL Pipeline, Data Cleaning, Data Validation, Data Analytics, Research and Development.

Contents

1	Introduction	1
1.1	Introduction	1
1.2	Objective of the internship report	2
2	Background	4
2.1	Overview of Cliqpack Limited	4
2.2	Mission of Cliqpack Limited	5
2.3	Vision of Cliqpack Limited	5
2.4	Industry Environment and Marketplace	5
2.5	Purpose of the Internship	7
3	Scope of Work	8
3.1	Automated Web Data Extraction and Information Retrieval	8
3.2	ETL Pipeline Maintenance and Data Integrity	9
3.3	Research and Development in CliqProperty	10
3.4	Project Management and Execution Planning	11
4	Internship Activities	12
4.1	Introduction of the Organization	12
4.2	Location and facilities of offices	12
4.3	Work Accomplishment	13
4.4	Challenges Faced	16
5	Methodology	18
5.1	Automated Web Scraping	18
5.2	ETL Pipeline Maintenance and Batch Processing	19
5.3	Data Validation	19

5.4	Research and Development of CliqProperty by using LLM Models .	20
6	Analysis and Findings	21
6.1	Results/Findings	21
6.2	Learning and Development	23
6.3	Tools and Technology	23
6.4	User Guideline	25
6.5	Impact of the Internship	25
7	Conclusions and Recommendations	27
7.1	Conclusion	27
7.2	Limitations	27
7.3	Recommendations	28
8	Acknowledgments	30
	References	33

List of Figures

2.1	Cliqpack Logo	4
4.1	Scraped data titled as “Naplan VCE Result with ICSEA Score ” . . .	14
4.2	Scraped data titled “School Profile”	14
4.3	Bangladesh-Data-Source-Inventory	14
4.4	Project Timeline for CliqProperty AI-Copilot	15
6.1	Scraped data titled as “Naplan-VCE-Result-with-ICSEA-Score” . . .	21
6.2	Figure: Scraped data titled as “School-Profile”	21
6.3	Figure: Bangladesh-Data-Source-Inventory	22
6.4	All Inner East Suburbs Agents’ Information	22
6.5	Prompting in Ollama	24
6.6	Prompting in Ollama	24
8.1	Appointment Letter	31
8.2	School Profile Website	32
8.3	Real Estate Website of Australia	32

List of Tables

Chapter 1

Introduction

1.1 Introduction

The internship program is a vital step in bridging the gap between the foundational knowledge of computer science and its practical application in the real world [Wahed et al. \(2024\)](#). Information has been the most important resource for corporate development during the last decade, with the internet serving as the primary source, and companies establish their business plans by obtaining and analyzing web data. It provided me with an opportunity to gain hands-on experience in a professional engineering environment at Cliqpack Limited. During these three months, I deepened my understanding of the end to end data scraping, including the extraction, processing, and strategic management of large scale datasets [Khder \(2021\)](#).

In this internship, I actively worked as a Data Science Engineering Intern, where my responsibilities centered on web scraping and the development of robust ETL (Extract, Transform, Load) pipelines. Web scraping is a technology that allows for the extraction of structured data from text such as HTML and is extremely useful in situations where data is not provided in a machine-readable format like JSON or XML [Thomas & Mathur \(2019\)](#). My tasks primarily involved using XPath and CSS selectors to extract structured data from complex platforms, such as [Australian Curriculum, Assessment and Reporting Authority \(ACARA\) \(2026\)](#) and Australian real estate directories like [REA Group \(2026\)](#).

I gained significant experience in helping construct automated pipelines to merge fragmented datasets specifically handling records for over 2,284 schools—and implementing rigorous validation checks to ensure data integrity. A critical part of

my work involved the introduction of automated validation methods to enhance the reliability of the datasets. By implementing checks for ID collisions and uniqueness between school IDs and years, I ensured that only accurate and non-duplicated information was stored in the master tables. This experience allowed me to see first-hand how effective data management and systematic optimization are essential for the performance of large-scale applications. I also contributed to the Research and Development of the “Cliqproperty” project, where I sourced public datasets from organizations like RAJUK, the World Bank, and the BBS to build a comprehensive property intelligence repository.

During the internship, I also encountered and solved real-world engineering problems, including handling dynamic web content, resolving data missingness, and managing anti-scraping mechanisms through polite crawling practices and rate limiting. While web scraping is gaining popularity, it also causes controversy due to ethical concerns regarding privacy and copyright [Wahed et al. \(2024\)](#).

These challenges significantly enhanced my problem-solving skills and technical competencies, particularly in Python-based data tools. Python is usually recognized as the best programming language for web scraping due to its ease of use and the availability of powerful libraries like BeautifulSoup and Selenium.

In addition, the professional environment at Cliqpack Limited promoted continuous learning and teamwork; I was able to communicate with experienced professionals, receive constructive feedback, and refine both my technical and communication skills

Altogether, I can state that the internship was a truly enriching experience that helped me grow personally and professionally. Web scraping as a data mining technology is gaining popularity among data scientists and analysts as it allows for the development of large, customized data sets at minimal costs [Wahed et al. \(2024\)](#).

It not only gave me the necessary experience and knowledge in the data science industry but also gave me the self confidence to become a professional Data Science Engineer.

1.2 Objective of the internship report

The following are the points that describe the objectives of my internship in Cliqpack:

- **Understand Data Science and Data Engineering Operations:** Acquire practical knowledge and develop a clear understanding of how data science and data engineering processes are implemented in a professional organizational environment.
- **Observe and Learn Data Processing Procedures:** Learn the procedures and responsibilities involved in data collection, data extraction, data cleaning, and data management within the Data Science Engineering Department, including website analysis, web crawling, and data organization processes. Develop Technical Skills in Data
- **Engineering Tools and Techniques:** Enhance understanding of web scraping techniques such as XPath and CSS selectors, data pipeline development (ETL), data preprocessing, and data validation methods.
- **Gain Practical Experience in Real-World Data Handling:** Observe and assist in handling real-world datasets, including identifying data quality issues such as missing values and duplicates, and applying appropriate data cleaning techniques.
- **Get Familiar with Professional and Corporate Work Culture:** Understand teamwork, communication standards, time management, and professionalism within a corporate data science and technology-driven environment.

Chapter 2

Background

2.1 Overview of Cliqpack Limited

Cliqpack Limited is a technology and operations company that provides data-driven digital solutions and IT services to businesses across various industries. The organization focuses on building intelligent software systems, managing digital operations, and supporting organizations in improving efficiency through data, automation, and modern technologies. Over the years, the company has established a reputation as a reliable technology partner for startups, enterprises, and institutions seeking scalable and innovative digital solutions. The company was founded in 2017 in Australia and later expanded its operations to Bangladesh in 2021, serving clients in multiple sectors such as real estate, education, retail, telecom, and agriculture.

Cliqpack Limited offers a wide range of services beyond traditional software development. These services include digital engineering solutions, web and mobile application development, data analytics and artificial intelligence (AI) solutions, cloud computing and cybersecurity services, data processing and automation systems, and business intelligence support. The organization also managed digital



Figure 2.1: Cliqpack Logo

operations and IT infrastructure services designed to improve productivity, accuracy, and operational efficiency for corporate clients.

2.2 Mission of Cliqpack Limited

The mission of Cliqpack Limited is to build data-driven, AI-powered, and people-led platforms that transform how industries and governments operate. Starting from emerging markets and scaling globally, the company focuses on empowering industries, institutions, and emerging economies through intelligent platforms and hybrid service delivery. By seamlessly bridging advanced technology with human expertise, Cliqpack drives sustainable growth, operational efficiency, and digital inclusion, while creating future-proof digital infrastructure designed to last for the next 50 years. This is what it takes into consideration to provide the best services for everyone.

2.3 Vision of Cliqpack Limited

The vision of Cliqpack Limited is to be the leading builder of vertical AI infrastructure that empowers key sectors such as real estate, education, agriculture, and public systems over the next 50 years. By shaping resilient and intelligent societies, the company aims to drive sustainable transformation across Australia, South Asia, MENA, and the Pacific regions. Cliqpack envisions creating future-proof digital infrastructure through intelligent platforms and hybrid service delivery, ultimately fostering growth, operational efficiency, and digital inclusion while bridging advanced technology with human expertise. Most companies increasingly demand for data from multiple websites to use in the development of their business [Wahed et al. \(2024\)](#).

2.4 Industry Environment and Marketplace

The global Property Technology and AI-powered data engineering industry is highly competitive and dynamic, driven by the rapid digital transformation across real estate management, market intelligence, educational data analytics, and AI advisory systems. Organizations are actively seeking customized ERP platforms,

automated web data extraction pipelines, large-scale dataset validation tools, and LLM-based copilots to achieve operational efficiency, data-driven decision-making, improved market accessibility, and sustainable competitive advantage in both developed and emerging markets.

Within this constantly evolving environment, CliqPack Ltd. has a unique market position by offering end-to-end AI-enabled solutions with a strong emphasis on real estate ERP systems and intelligent data infrastructure.

There are several strong points of the market positioning of the company:

- **Highly Differentiated Services:** CliqPack provides a complete portfolio that spans AI-powered real estate ERP (CliqProperty), large-scale educational and financial dataset scraping and validation, and custom AI Copilot development. This flexibility enables the company to serve multiple high-value verticals, including real estate, education, and government sectors, with tailored, production-ready solutions.
- **Technical Proficiency:** The company demonstrates deep expertise in modern data engineering stacks, including XPath/CSS selector-based web scraping, dataset merging and validation, local LLM deployment (Ollama + Llama 3.1), prompt engineering, refusal design, and compliant lead-generation pipelines. This technical strength allows CliqPack to deliver accurate, scalable, and ethically sound data assets for complex business requirements.
- **High Customer Base:** CliqPack's growing client portfolio across Australia and Bangladesh, covering real estate developers, educational institutions, and government-linked projects, confirms the company's capability to handle large-scale, cross-border data and AI initiatives for demanding enterprise clients.
- **Global Delivery Model:** Headquartered in Melbourne, Australia, with development and operations centers in Dhaka and Chattogram, Bangladesh, CliqPack offers a cost-effective global delivery model that combines local project management, regulatory compliance knowledge, and international market insight, providing superior value to clients in both mature and emerging markets.

2.5 Purpose of the Internship

The internship at Cliqpack Limited was designed to provide hands-on experience to students who are relatively new to the field of Data Science Engineering. It aimed to bridge the gap between academic knowledge and real-world industry practices by allowing interns to work alongside experienced engineers in a professional technology environment.

The primary purpose was to gain practical exposure in data-related workflows within an AI-powered software and service company. This included understanding how data is extracted, processed, and utilized to support intelligent platforms. The internship offered an excellent opportunity to move beyond traditional classroom learning and apply theoretical concepts of data science, programming, and data management to solve actual business and technical problems.

Another important objective was to develop a strong foundational understanding of data engineering tasks commonly performed in technology organizations. During the internship, I focused on learning web scraping techniques using XPath and CSS selectors, building and managing ETL data pipelines, data cleaning, preprocessing, validation, and merging of datasets from multiple sources. These activities helped me understand how high-quality data is prepared to power AI systems and business intelligence solutions. Information has been the most important resource for corporate development during the last decade, with the internet serving as the primary source, and companies establish their business plans by obtaining and analyzing web data [Wahed et al. \(2024\)](#).

Additionally, the internship played a significant role in developing essential professional skills such as problem-solving, attention to detail, time management, and teamwork. Working closely with skilled data engineers and technical teams enhanced my ability to manage real-world data challenges and improved my communication skills in a corporate setting.

Overall, this internship helped me to understand how to build a career in Data Science and Data Engineering by providing practical insight into how modern technology companies manage data to deliver intelligent and scalable services to their clients.

Chapter 3

Scope of Work

The scope of my internship at Cliqpack Limited was defined by three primary technical domains: high-precision web data extraction, the maintenance of robust ETL (Extract, Transform, Load) pipelines, and strategic research and development for AI-integrated platforms. My responsibilities were designed to bridge the gap between raw web data and commercial-grade intelligence. Web scraping or web crawling refers to the procedure of automatic extraction of data from websites using software [Khder \(2021\)](#). The specific tasks and projects are detailed below:

3.1 Automated Web Data Extraction and Information Retrieval

Website analysis, website crawling, and data organizing are the three primary, interwoven processes of online scraping [Khder \(2021\)](#). During the internship, I designed and implemented automated web scraping pipelines for large-scale structured data extraction from multiple public platforms, including [myschool.edu.au](#), [realestate.com.au](#), and [domain.com.au](#). I utilized XPath and CSS selectors for precise information retrieval of school profiles, NAPLAN/VCE results, financial metrics, as well as real estate agent and agency lead data. To ensure reliability and ethical compliance, I developed batch processing mechanisms with polite crawling, rate limiting, caching, and retry logic. Finally, I performed systematic data merging, normalization, deduplication, and validation to transform raw web data into clean, production-ready datasets that directly support CliqProperty AI Copilot development and CliqPack's Australian market expansion.

A significant portion of data extraction involves the retrieval of structured data from complex, multi-layered web environments using XPath and CSS selectors. I was responsible for:

- **Educational Data Sourcing:** Targeted extraction from the Australian School Directory (myschool.edu.au), capturing school sectors, location data, and staff statistics. Retrieving specialized educational indicators, including the Index of Community Socio-Educational Advantage (ICSEA) and NAPLAN VCE results.
- **Real Estate Directory Mapping:** Automating the collection of Australian real estate agent and agency leads from platforms like realestate.com.au and domain.com.au, focusing on the hierarchical structure of State → Suburb → Agency → Agent.

3.2 ETL Pipeline Maintenance and Data Integrity

In addition to extraction, my role was to manage the data engineering lifecycle to ensure that the resulting datasets were production-ready. During the internship, I designed and implemented robust ETL (Extract, Transform, Load) pipelines to process and integrate large-scale datasets from multiple fragmented sources for approximately 2,284 Australian schools. Web scraping is a technique for converting unstructured web data into structured data that can be saved and analyzed in a central spreadsheet or database [Khder \(2021\)](#). Using Python and Pandas, I performed data merging, cleaning, normalization, and transformation while enforcing strict data integrity rules, including ID collision checks, composite-key uniqueness validation, and range constraints on percentage-based columns. Automated validation scripts were executed immediately after every merge, followed by systematic manual cross-verification against the live website to guarantee 100% accuracy and completeness. This work transformed raw, batched data into clean, reliable, production-ready datasets that serve as a solid foundation for CliqProperty AI Copilot analytics and CliqPack's educational data platform initiatives.

This included:

- **Dataset Merging:** Developing scripts to merge fragmented datasets such as the batch-scraped records for 2,284 schools into consolidated main tables.

- **Automated Validation Frameworks:** Implementing rigorous “must-have” validation checks to prevent ID collisions, ensure the uniqueness of (school-id, year) pairings, and maintain the integrity of percentage-based columns.
- **Normalization and Deduplication:** Cleaning and standardizing agency names and agent records to create stable, “lead-ready” exports for CRM integration into systems.

3.3 Research and Development in CliqProperty

My task was to conduct the foundational research for the CliqProperty AI Co-pilot, an LLM-powered advisory tool. During the internship, I conducted a comprehensive Strategic R and D for the CliqProperty AI Copilot, a new LLM-powered advisory system for real estate operations. I thoroughly studied the Co-pilot doctrine, platform boundaries, ecosystem map, and strategic focus areas, followed by systematic identification and evaluation of publicly available real estate datasets from authoritative Bangladeshi sources, including BBS, RAJUK, land.gov.bd, Bproperty, OSM, and international bodies such as the World Bank and ADB. Utilizing machine learning could result in scraping techniques that automatically “learn” from experience and can adapt to a wider variety of sites [Khder \(2021\)](#). These datasets were analyzed and organized into three strategic layers: Market Intelligence, Operational, and Spatial/Government data to ensure seamless integration with the platform. I also designed and delivered a detailed 10-month (200 working days) execution timeline covering local LLM setup (Ollama + Llama 3.1 8B), prompt engineering, refusal taxonomy, and integration with <https://cliqproperty.io/>, completing the initial phase up to Day 10. This strategic foundation enables the development of a secure, read-only AI Copilot that will significantly enhance CliqPack’s intelligent real estate offerings in Bangladesh.

My scope in this area involved:

- **Public Dataset Sourcing:** Identifying and categorizing over 8,000 property listings and governmental datasets from sources including RAJUK, the World Bank, Asian Development Bank (ADB), and the BBS.

- **LLM Benchmarking and Implementation:** Evaluating the performance of local LLMs, specifically Llama 3.1 8B (Q4 GGUF), using the Ollama runtime to determine latency and CPU threading behavior.
- **Security and Policy Design:** Defining the refusal taxonomy and doctrine for the AI Copilot to mitigate risks such as prompt injection and unauthorized “ledger writes” in financial contexts.

3.4 Project Management and Execution Planning

During the internship, I led the Project Management and Execution Planning for the CliqProperty AI Copilot v1 initiative by designing a comprehensive 10-month (200 working days) execution timeline with detailed day-wise schedules, resource allocation, and milestone deliverables. I developed a structured plan covering local LLM setup (Ollama + Llama 3.1 8B Q4 GGUF), prompt control, refusal taxonomy, and integration with <https://cliqproperty.io/>, while clearly defining weekly objectives, risk considerations, and technology decisions. After thorough review and approval from my supervisor, I successfully executed the initial phase up to Day 10, maintaining strict adherence to the defined timeframe and documentation standards. This proactive project management approach ensured organized, measurable progress and provided a clear roadmap for the full development of the AI Copilot system. Finally, my role required me to operate within professional project management frameworks by:

- **Timeline Drafting:** Constructing a detailed 10-month execution schedule (200 working days) for the AI Copilot project, outlining milestones for system readiness, prompt control, and deployment.
- **Compliance Monitoring:** Ensuring all scraping activities maintain the polite crawling standards. The privacy requirements for contact data need to ensure polite crawling.

Chapter 4

Internship Activities

4.1 Introduction of the Organization

My internship was conducted at Cliqpack Limited, a technology-oriented organization specializing in high-performance data engineering. The organization is deeply involved in building scalable infrastructures for data acquisition and market intelligence, serving international markets including Australia and Bangladesh. Web scraping or web crawling refers to the procedure of automatic extraction of data from websites using software, a process particularly important in fields such as Business Intelligence in the modern age [Khder \(2021\)](#). During my attachment, the engineering department focused on developing CliqProperty, a platform that integrates local and spatial datasets with large language models (LLMs) to provide safe property advice. Information has been the most important resource for corporate development during the last decade, with companies establishing business plans by obtaining and analyzing web data [Wahed et al. \(2024\)](#).

4.2 Location and facilities of offices

CliqProperty is a flagship AI-powered real estate ERP platform developed and operated by CliqPack Ltd. It maintains a strong global presence with its headquarters in Melbourne, Australia, and major development and operational facilities in Dhaka and Chattogram, Bangladesh.

The Australian headquarters is located at 55 Collins Street, Suite 147 / Level 1, Melbourne, Victoria 3000, providing strategic oversight, business development, and international client coordination. In Bangladesh, the Dhaka office is situated

at Green Trade Point,7, C/A, Bir Uttam AK Khandakar Road, Mohakhali, Dhaka-1212. The Chattogram office is based at R.I Tower (3rd Floor), 23/A M M Ali Road, Golpahar Circle, Mehedibag, Chattogram 4000.

Service Day	Service Hours
Sunday – Thursday	9:00 – 6:00
Friday	Closed
Saturday	Closed

These modern, well-equipped facilities house dedicated teams for software engineering, data infrastructure, AI development, and customer support, enabling 24/7 operational efficiency and seamless delivery of the CliqProperty platform to clients across both markets.

4.3 Work Accomplishment

The following is a summary of the activities performed during my three-month tenure, structured to highlight the transition from theoretical learning to industrial application.

Month 1: Data Extraction and Quality Assurance (The Australian Schools Project)

- **Observed:** I observed the internal protocols for inspecting website DOM structures to identify reliable data hooks.
- **Performed:** I inspected and extracted structured data from myschool.edu.au using XPath and CSS selectors. This included school demographics, ICSEA values, staff statistics, and financial recurrent funding data. Web scraping is a technology that allows us to extract structured data from text such as HTML and is extremely useful when data is not provided in machine-readable formats. [Khder \(2021\)](#).
- **Learned:** I developed an automated pipeline to merge datasets for 2,284 schools. Because the source website prevented bulk scraping, I learned to process data in batches of 10 school IDs. It has been found that using a web scraping program yields data that is far more thorough, accurate, and consistent than manual entry [Khder \(2021\)](#).

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
1	id_x	school_id	year	grade_level_id	year_range	school_icse_value	year_3_reading_score	year_3_writing_score	year_3_pelling_score	year_3_grammar_score	year_3_numeracy_score	year_5_reading_score	year_5_writing_score	year_5_pelling_score	year_5_grammar_score	year_5_numeracy_score	year_7_reading_score	year_7_writing_score	year_7_pelling_score	year_7_grammar_score	year_7_numeracy_score	year_9_reading_score	year_9_writing_score
593	10152	44148	2025	Primary	Prep-6	974	403	424	383	395	402	503	452	463	503	488							
594	563	44149	2024	Combined	Prep-10	934	526	574	561	524	538	467	431	463	491	503	527	497	515	498	-	530	573
595	564	44149	2023	Combined	Prep-10	961	548	589	563	570	425	450	418	437	455	448	-	-	-	-	-	-	-
596	565	44149	2022	Combined	Prep-10	941	450	384	404	422	441	526	484	504	498	510	533	537	535	510	524	561	604
597	566	44149	2021	Combined	Prep-10	976	407	409	392	441	445	506	460	476	506	484	511	494	485	478	498	-	-

Figure 4.1: Scraped data titled as “Naplan VCE Result with ICSEA Score ”

	ACARA School ID	Year	School sector	School Type	Year range	Location	Teaching staff	Non-teaching staff	Full-time equivalent teaching staff	Full-time equivalent non-teaching staff	School website (text)	School website (href)	Sector/system /association website (text)	Sector/system /association website (href)	School sub-campus (text)	School value	Average ICSE A value	School percentile	Bottom Quartile Scholastic Distri	Bottom Quartile Scholastic Distri	1st Middle Quartile Scholastic Distri	1st Middle Quartile Scholastic Distri	2nd Middle Quartile Scholastic Distri	2nd Middle Quartile Scholastic Distri	Top quartile Scholastic Distri	Top quartile Scholastic Distri	SEA legend note	Total enrolments
1																												
2	44524	2025	Governme	Primary	Prep-6	Major Citi	24	7	19.2	5.2	Moonee P	http://www.mooniep.shs.vic.edu.au/	Department of	https://www.y.schools.vic.edu.au/	School fac	1136	1000	92	5%	25%	14%	25%	28%	25%	55%	25%	Percentage	Total enro
3	44525	2025	Governme	Primary	Prep-6	Major Citi	40	21	32.5	11.9	Mitcham P	http://www.mitcham.p.shs.vic.edu.au/	Department of	https://www.y.schools.vic.edu.au/	School fac	1127	1000	91	6%	25%	13%	25%	30%	25%	51%	25%	Percentage	Total enro
4	44526	2025	Governme	Primary	Prep-6	Inner Regi	16	14	11.4	9.4	Loch Prim	http://www.lochprim.shs.vic.edu.au/	Department of	https://www.y.schools.vic.edu.au/	School fac	995	1000	45	37%	25%	37%	25%	20%	25%	6%	25%	Percentage	Total enro
5	44527	2025	Governme	Primary	Prep-6	Outer Regi	33	39	29.1	31.4	Mildura Pr	http://www.mildura.p.shs.vic.edu.au/	Department of	https://www.y.schools.vic.edu.au/	School fac	797	1000	3	84%	25%	11%	25%	4%	25%	1%	25%	Percentage	Total enro

Figure 4.2: Scraped data titled “School Profile”

- **Theory vs. Practice Gap:** In academic projects, datasets are often provided in a clean, unified format. In industry, I found that data acquisition is highly fragmented and requires rigorous post-concat integrity checks to prevent ID collisions and ensure the uniqueness of (school-id, year) pairings.

Month 2: Research, Development, and AI Planning (The Cliqproperty Project): I experienced the Research and Development lifecycle for a new product, Cliqproperty AI Co-pilot.

- **Performed:** I sourced over 8,000 public property listings and categorized them into three layers: Market Intelligence, Operational Data, and Spatial/Government Data. I also drafted a 10-month execution schedule for the AI Copilot project. Data acquisition is recognized as the first phase in any data science study and development **Khder (2021)**.

1	Category	Source	Dataset / Asset	Format	Update Frequency	Access	Link	Primary Use in CliqProperty	Data Source's Path
2	Official Statistics	BBS NSDS	Population & Housing Census 2022 (overview/posts)	Web/PDF	Periodic	Public	http://203.112.2	Census context for household/stock normalization and locality profiling.	Population and Housing Census 2022.pdf
3	Land Records	Public Works Department (PWD)	Land Information reports on acquired land and illegally occupied land	Web-based reports	Ongoing	Public	https://land.gov.d.gov.bd/view-report.php	Can give details of land and buildings under the control of the Ministry of Housing and Public Works and its subordinate directorates/departments/autonomous bodies/divisions/district and upazilla offices	Khatian.png
4	Land Records	land.gov.bd	Bangladesh Land Information Portal	Web	Ongoing	Public	https://land.gov	Official gateway to land-related services and information.	
5	Land Records	DLRMS (Land)	DLRMS portal (land record services)	Web	Ongoing	Public	https://dlrms.lar	Land record references.Can get details of Survey Khatyan.	Khatian (1).png
6	Geospatial	RAJUK	RAJUK DAP downloads	Web/PDF	Periodic	Public	https://rajuk.gov	Can be used to get detailed area plan.	RAJUK DAP.pdf
7	Geospatial	RAJUK GIS Portal	RAJUK GIS portal entry	Web	Varies	Public/Restrictet	https://wplcs.raj	A Geographic Information System that can give geographical information.	Gulshan-Central-East-North.jpg
8	Market Data	Bproperty	Property listings (sale/rent), market insights blogs	Web	Ongoing	Public	https://bproperty.com/apartments/	Real-time property listings and price,rent benchmarks in Dhaka	

Figure 4.3: Bangladesh-Data-Source-Inventory

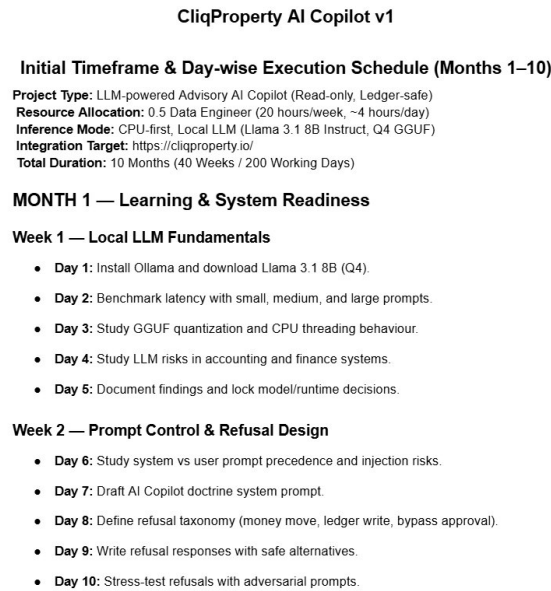


Figure 4.4: Project Timeline for CliqProperty AI-Copilot

I performed local LLM benchmarking using Ollama and Llama 3.1 8B (Q4 GGUF). I understood how CPU threading and GGUF quantization affect latency in a production environment.

- **Theory vs. Practice Gap:** While academic research often focuses on model accuracy, industrial Research and Development at Cliqpack prioritized refusal taxonomy, which designs system prompts to prevent ledger writes or unauthorized financial advice.

Month 3: Australian Real Estate data scraped:

- **Performed:** I helped to build a dataset for Australian agencies and agents using realestate.com.au and domain.com.au. This involved creating a mapping table (agency-agent-map.csv) and segmenting leads by state and role. I developed a pilot dataset for Victoria (VIC), focusing on growth corridors to enable immediate outbound marketing. Real property list collecting is one of many scenarios that web scraping may be utilized for [Wahed et al. \(2024\)](#).
- **Learned:** I learned to implement polite crawling practices, including rate limiting and respecting robots.txt guidance to maintain legal compliance.

- **Theory vs. Practice Gap:** Academic scraping often ignores the ethical and legal constraints of data collection. In practice, managing privacy requirements and "last verified" timestamps is mandatory for commercial-grade databases.

4.4 Challenges Faced

During my internship, I faced several technical and practical problems. I learned a lot while solving them with simple and smart solutions.

- **Changing Website Designs:** Websites such as myschool.edu.au, realestate.com.au, and domain.com.au often changed their page layouts. Because of this, my XPath selectors stopped working, and I got wrong or missing data. A significant change in the targeted site's design or behavior might make it extremely difficult for your program to find the appropriate data on the site [Khder \(2021\)](#). I fixed this by using stronger CSS selectors and adding a quick check script that flagged the issues right away.
- **Lack of Proper Resources:** When I tested the Llama 3.1 AI model on my normal computer (CPU only), it was very slow and took too much time. I solved this by studying the model settings, trying different quantization levels, and writing notes on how to make it faster.
- **Website Blocking Scrapers:** When I tried to collect thousands of school and agency records, the websites blocked me to stop automatic scraping. When online scraping procedures scale, many challenges may arise, such as blocking mechanisms which can prevent scrapers from receiving data. I handled this by adding slow and careful crawling, small batch sizes, waiting times, and caching so it looked like normal human browsing.
- **Too Much Data to Combine:** I had to merge data from 2,284 schools that came in small batches. Sometimes old and new data are mixed up or had missing parts. I created simple checking rules to stop collisions and double-check everything before combining.
- **Time-Consuming Manual Checking:** After the automatic checks, I still had to manually compare many records with the original website. This took a

long time. I solved it by picking random samples and working closely with my supervisor whenever I found mistakes.

- **Finding Good Datasets for Bangladesh:** It was hard to find complete and up-to-date real estate data for CliqProperty AI Copilot. Many sources had old information or missing details. I searched many government websites and carefully selected the best ones.

These challenges helped me become better at solving real problems. I learned how to work patiently, think step by step, and always keep data clean and correct.

Chapter 5

Methodology

The methodology adopted during this internship at Cliqpack Limited followed a systematic data engineering lifecycle, transitioning from raw data acquisition to high-integrity validation and strategic AI research. My approach was divided into four primary technical phases:

5.1 Automated Web Scraping

The initial phase involved the targeted extraction of structured data from complex web environments using XPath and CSS selectors. Understanding the HTML structure of a website is crucial for successful internet scraping as it allows scrapers to correctly target and extract the preferred data using tags like `<div>`, `<span>`, and `<a>` [Wahed et al. \(2024\)](#)

- **Targeted Information Retrieval:** For the Australian school project, I extracted multiple data points, including school sectors, staff statistics, and the Index of Community Socio-Educational Advantage (ICSEA).
- **Polite Crawling Protocols:** To maintain ethical compliance and avoid IP blocking, I implemented polite crawling techniques. For high-volume directories like `realestate.com.au`, data was captured in a structured pipeline of `State → Suburb → Agency → Agent`.

5.2 ETL Pipeline Maintenance and Batch Processing

Following the extraction, I developed Extract, Transform, Load (ETL) pipelines to manage and consolidate fragmented datasets. Web scraping is a technique for converting unstructured web data into structured data that can be saved and analyzed in a central spreadsheet or database [Khder \(2021\)](#).

- **Batch Scraping:** Because bulk scraping was restricted, I developed a script to process data in batches of 10 school IDs, eventually merging records for 2,284 schools into a master table.
- **Data Normalization:** For the agency lead database, I implemented normalization rules to standardize company names by removing punctuation and legal suffixes like "Pty Ltd". I also generated stable IDs (agent-id and agency-id) to ensure long-term traceability.

5.3 Data Validation

To ensure the resulting datasets were safe, I implemented a multi-layered validation framework immediately after the data concatenation process. This aligns with the principles of "Tidy Data," where structural consistency is paramount (Wickham, 2014). Proper data preprocessing and cleanup are required for the proper utilization of scraped datasets [Khder \(2021\)](#).

I developed automated scripts to enforce three "must-have" validation rules:

- **ID Collision Verification:** Ensuring the intersection of the old and new ID sets remained empty.
- **Temporal Uniqueness:** Confirming that (school-id, year) pairings were unique across the combined dataset.
- **Boundary Validation:** Ensuring that all percentage-based columns (e.g., participation or funding metrics) remained strictly within the 0–100 range.

Manual Cross-Checking: I performed randomized manual validation by sampling output records and verifying them directly against the source website UI. If discrepancies were found, I modified the scraping logic under supervisory guidance to achieve the desired accuracy.

5.4 Research and Development of CliqProperty by using LLM Models

For the "Cliqproperty" project, the methodology shifted toward secondary research and technical feasibility testing for AI integration.

- **Public Dataset Sourcing:** I conducted a qualitative search for public assets, sourcing data from the World Bank, RAJUK, BBS, and the Asian Development Bank (ADB). These were categorized into three functional layers: Market Intelligence, Operational Data, and Spatial/Government Data.
- **Inference Benchmarking:** I utilized Ollama to install and benchmark Llama 3.1 8B (Q4 GGUF). My methodology involved measuring latency across small, medium, and large prompts while monitoring CPU threading behavior to optimize local deployment (Touvron et al., 2023).
- **Refusal Taxonomy Design:** I drafted an AI Copilot doctrine that defined a refusal taxonomy, identifying specific risks like "ledger writes" or "money moves" to ensure the AI remained a "read-only" advisory tool. This included adversarial stress-testing to verify the robustness of the system prompts against injection risks.

Chapter 6

Analysis and Findings

6.1 Results/Findings

The internship at Cliqpack Limited resulted in the successful delivery of three major data engineering assets, each dataset validated for commercial use.

- **Validated Educational Dataset:** I successfully processed and merged structured data for 2,284 schools from the MySchool directory. This final dataset includes high-fidelity information on ICSEA values, socio-educational advantage (SEA) distributions, and student enrollment demographics. Furthermore, the repository contains detailed NAPLAN VCE results and financial recurrent funding data, encompassing government contributions and parent fees per student.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
1	id_x	school_id	year	grade_level_id	year_range	school_icsea_value	year_3_reading_score	year_3_writing_score	year_3_pelling_score	year_3_grammar_score	year_3_umeracy_score	year_5_reading_score	year_5_writing_score	year_5_pelling_score	year_5_grammar_score	year_5_umeracy_score	year_7_reading_score	year_7_writing_score	year_7_pelling_score	year_7_grammar_score	year_7_umeracy_score	year_9_reading_score	year_9_writing_score
593	10152	44148	2025	Primary	Prep-6	974	403	424	383	395	402	503	452	463	503	488	-	-	-	-	-	-	-
594	563	44149	2024	Combined	Prep-10	934	326	374	361	324	358	467	431	463	491	503	527	497	515	498	-	530	573
595	564	44149	2023	Combined	Prep-10	961	348	389	363	370	425	450	418	437	455	448	-	-	-	-	-	-	-
596	565	44149	2022	Combined	Prep-10	941	450	384	404	422	441	526	484	504	498	510	533	537	535	510	524	561	604
597	566	44149	2021	Combined	Prep-10	976	407	409	392	441	445	506	460	476	506	484	511	494	485	478	498	-	-

Figure 6.1: Scraped data titled as “Naplan-VCE-Result-with-ICSEA-Score”

	ACARA School ID	Year	School sector	School Type	Year range	Location	Teaching staff	Non-teaching staff	Full-time equivalent teaching staff	Full-time equivalent non-teaching staff	School website (text)	School website (href)	Sector/system /association website (text)	Sector/system/association website (href)	School sub-campus (text)	School ICSEA value	Average ICSEA A value	School ICSEA A percentile	Bottom Quartile School allian Distr	Bottom Quartile School allian Distr	1st Middle Quartile School allian Distr	1st Middle Quartile School allian Distr	2nd Middle Quartile School allian Distr	2nd Middle Quartile School allian Distr	Top quartile School allian Distr	Top quartile School allian Distr	SEA legend note	Total enrolments
1																												
2	44524	2025	Govermme	Primary	Prep-6	Major City	24	7	19.2	5.2	Moonee P	http://www.mooniep.com.au	Department of	https://www.vic.gov.au	School fac	1136	1000	92	3%	25%	14%	25%	28%	25%	55%	25%	Percentage	Total enro
3	44525	2025	Govermme	Primary	Prep-6	Major City	40	21	22.5	11.9	Mitcham P	http://www.mitchamp.com.au	Department of	https://www.vic.gov.au	School fac	1127	1000	51	6%	25%	13%	25%	30%	25%	51%	25%	Percentage	Total enro
4	44526	2025	Govermme	Primary	Prep-6	Inner Regi	16	14	11.4	9.4	Loch Prim	http://www.lochprim.com.au	Department of	https://www.vic.gov.au	School fac	995	1000	45	57%	25%	57%	25%	20%	23%	6%	25%	Percentage	Total enro
5	44527	2025	Govermme	Primary	Prep-6	Outer Regi	33	39	29.1	31.4	Mildura Pr	http://www.mildura.com.au	Department of	https://www.vic.gov.au	School fac	797	1000	3	84%	25%	11%	25%	4%	25%	1%	25%	Percentage	Total enro

Figure 6.2: Figure: Scraped data titled as “School-Profile”

- **Property Intelligence Repository:** For the “Cliqproperty” project, I identified and categorized critical datasets into three operational layers: Market Intelligence, Operational Data, and Spatial/Government Data. This included sourcing over 8,000 property listings across major cities in Bangladesh, detailing property types, sizes, and neighborhood price trends.

1	Category	Source	Dataset / Asset	Format	Update Frequency	Access	Link	Primary Use in CliqProperty	Data Source's Path
2	Official Statistics	BBS NSDS	Population & Housing Census 2022 (overview/posts)	Web/PDF	Periodic	Public	http://203.112.2	Census context for household/stock normalization and locality profiling.	Population and Housing Census 2022.pdf
3	Land Records	Public Works Department (PWD)	Land Information reports on acquired land and illegally occupied land	Web-based reports	Ongoing	Public	https://land.gov.bd/view_report.php	Can give details of land and buildings under the control of the Ministry of Housing and Public Works and its subordinate directorates/departments/autonomous bodies/divisions/district and upazila offices	Khatian.png
4	Land Records	land.gov.bd	Bangladesh Land Information Portal	Web	Ongoing	Public	https://land.gov	Official gateway to land-related services and information.	
5	Land Records	DLRMS (Land)	DLRMS portal (land record services)	Web	Ongoing	Public	https://dlrms.lag.khatyan	Land record references.Can get details of Survey	Khatian (1).png
6	Geospatial	RAJUK	RAJUK DAP downloads	Web/PDF	Periodic	Public	https://rajuk.gov	Can be used to get detailed area plan.	RAJUK DAP.pdf
7	Geospatial	RAJUK GIS Portal	RAJUK GIS portal entry	Web	Varies	Public/Restricted	https://gis.rajuk.gov	A Geographic Information System that can give geographical information.	Gulshan-Central-East-North.jpg
8	Market Data	Bproperty	Property listings (sale/rent), market insights blogs	Web	Ongoing	Public	https://bproperty.com/apartments/	Real-time property listings and price,rent benchmarks in Dhaka	

Figure 6.3: Figure: Bangladesh-Data-Source-Inventory

- **Australian Real Estate Dataset:** I delivered a structured dataset for the Australian real estate market, featuring tables for agents and agencies. The web contains lots of interesting data sources that provide a treasure trove for all sorts of interesting things. Sadly, the current unstructured nature of the web does not always make it easy to gather or export this data in an easy manner.[vanden Broucke & Baesens \(2017\)](#) A key finding was the successful mapping of agencies to agents (agency-agent-map.csv), specifically for a pilot run in Victoria (VIC), which targeted growth corridors and inner-south suburbs to enable immediate marketing outbound.

	suburb	state	postcode	global_agent_number	page_number	agency_name	full_name	title_role	agency_name	rating_star	avg_sold_price	avg_days_on_market	prop_erties_sold	full_name	title_role	agency_name	office_address	phone	total_sales	average_sales_price	total_sales_value	average_days_on_market	profile_url	agent_link	agency_url	suburb_url	status	source_file
2	BERWICK	VIC	3806	1	1	Hayley Tai Partner & Barry Plan 5.0 StarOS 5966k	Hayley Tai Sales repri Barry Plan 100 High	1	+6140009 109	5943k	\$89.6m	22	https://w	https://w	https://w	https://w	https://w	Success	all_agents									
3	BERWICK	VIC	3806	2	1	Debbie Bn Director Ray White 4.9 StarOS 5977k	Debbie Bn Sales repri Ray White 42 High St	42	+6145862 65	\$1.0m	\$65.1m	34	https://w	https://w	https://w	https://w	https://w	Success	all_agents									
4	BERWICK	VIC	3806	3	1	Kristen Tu Director Peake Real Estate \$1.4m	Kristen Tu Principal Peake Rea 18 Langm	18	+6141061 70	\$1.3m	\$92.1m	44	https://w	https://w	https://w	https://w	https://w	Success	all_agents									
5	BERWICK	VIC	3806	4	1	Sam Noor Partner Ray White 5.0 StarOS \$1.0m	Sam Noor Sales repri Ray White 42 High St	42	+6141195 65	\$921k	\$59.0m	31	https://w	https://w	https://w	https://w	https://w	Success	all_agents									
6	BERWICK	VIC	3806	5	1	Ryan Wall Director JR Property \$1.3m	Ryan Wall Principal JR Property 2/7 Glouc	2/7	+6143262 71	\$2.4m	\$149.6m	60	https://w	https://w	https://w	https://w	https://w	Success	all_agents									
7	BERWICK	VIC	3806	6	1	Cameron I Licensed E Peake Real Estate \$1.0m	Cameron I Sales repri Peake Rea 18 Langm	18	+6142876 29	\$1.0m	\$29.1m	38	https://w	https://w	https://w	https://w	https://w	Success	all_agents									
8	BERWICK	VIC	3806	7	1	Daniel Boy Licensed E OBrien Real Estate Bk \$983k	Daniel Boy Sales repri OBrien Re. 56-58 Hlg	56-58	+6148878 39	\$987k	\$30.6m	33	https://w	https://w	https://w	https://w	https://w	Success	all_agents									
9	BERWICK	VIC	3806	8	1	Julie Petty Sales Cons Barry Plan 4.8 StarOS \$1.1m	Julie Petty Sales repri Barry Plan 100 High	1	+6147755 28	\$1.1m	\$28.8m	56	https://w	https://w	https://w	https://w	https://w	Success	all_agents									
10	BERWICK	VIC	3806	9	1	Paul Roge Licensed E OBrien Real Estate Bk \$1.0m	Paul Roge Sales repri OBrien Re. 56-58 Hlg	56-58	+6141856 30	\$1.0m	\$26.0m	23	https://w	https://w	https://w	https://w	https://w	Success	all_agents									

Figure 6.4: All Inner East Suburbs Agents' Information

- **CliqProperty AI-Copilot Implementation Timeframe:** I established a Implementation Timeframe for the CliqProperty AI Co-pilot, featuring a 10-month execution schedule that outlines milestones from local LLM installation to adversarial stress-testing of refusal responses.

6.2 Learning and Development

Throughout the internship, I gained a comprehensive understanding of the industrial data engineering lifecycle. Web scraping is a highly useful tool in the information age, and an essential one for any company wishing to keep up in the 21st century and maintain a strong online presence. [Thomas & Mathur \(2019\)](#).

- **Technical Proficiency:** I mastered the use of XPath and CSS selectors for complex web scraping and learned to build automated pipelines that process data in stable batches to avoid connection failures. Python is usually recognized as the best programming language for web scraping due to its ease of use and the availability of powerful libraries like BeautifulSoup, and Selenium.
- **Data Validation:** I developed the discipline of running validation scripts, such as checking for ID collisions and ensuring (school-id, year) uniqueness, which are critical for maintaining safe databases.
- **AI and Prompt Engineering:** I gained hands-on experience with Ollama and Llama 3.1 8B, learning how GGUF quantization and CPU threading behavior impact latency. I also learned the importance of refusal taxonomy, designing system prompts that prevent unauthorized financial operations.
- **Professional Growth:** My analytical skills were enhanced through regular supervisor feedback sessions, where I learned to modify scraping logic based on real-world discrepancies found during manual cross-checks.

6.3 Tools and Technology

During the internship, I used a set of simple and powerful tools and technologies to complete all my tasks efficiently. Extracting quality data from the web requires a combination of robust programming environments and specialized data extraction libraries. [Chapagain \(2023\)](#) These tools helped me with coding, data scraping, AI testing, data checking, and report writing.

- **Visual Studio Code (VS Code):** My main code editor for writing and running Python scripts. It made coding, debugging, and organizing files very easy.
- **ChromeDriver with Selenium:** I used this for automated web scraping. It helped me open websites, find data using XPath and CSS selectors, and collect information from myschool.edu.au, realestate.com.au, and domain.com.au.
- **Ollama:** I used Ollama to run the Llama 3.1 8B LLM model locally on my computer. It was very useful for testing the AI Copilot, checking prompt responses, and doing benchmarking.

```
>>> Can rent be paid using bKash in Bangladesh?
Yes, rent can be paid using bKash in Bangladesh. Many landlords and
property owners accept bKash as a mode of payment. This is because
bKash offers a
convenient and secure way for tenants to pay their rent online, wit
hout the need to physically visit the landlord's location.

To pay rent using bKash, tenants can use the following steps:

1. Open the bKash app or log in to the website.
2. Go to the "Pay" section and select the "Rent Payment" option.
3. Enter the payment details, including the amount of rent and the
landlord's account information.
4. Confirm the payment and make the transaction.

Once the payment is processed, tenants will receive a confirmation
message on their phone, and the landlord will also be notified.

total duration:      1m5.0784796s
load duration:       466.1975ms
prompt eval count:    136 token(s)
prompt eval duration: 8.2885052s
prompt eval rate:     16.41 tokens/s
eval count:          165 token(s)
eval duration:       55.4189537s
eval rate:           2.98 tokens/s
```

Figure 6.5: Prompting in Ollama

```
>>> You are a helpful property accounting assistant for CligProperty.
... Explain the tenant's balance in simple words.
... Use only this information:
...
... Flat: Mirpur-10, Dhaka
... Tenant: Rahim Mia
... Monthly rent: 12,000 Taka
... Last payment: 5 January 2026 - 12,000 Taka (paid by bKash)
... February rent: Not paid yet
... Water bill: 800 Taka (added in February)
... Previous balance: 0 Taka
...
... What is the current balance? Why is it like this? What would you tell the tenant?
The tenant's balance is -1,050 Taka. The balance has become negative because the tenant was supposed to p
ay the rent for February but didn't do so
yet.

You could explain this to the tenant by saying: "Dear Rahim Mia,
Your current balance is -1,050 Taka because you haven't paid the rent for February and also the water bil
l that was added in February. Your previous
balance was 0 Taka, and then on January 5, 2026, you paid 12,000 Taka as your monthly rent which brought
it to a positive balance. However, when the
new month starts (February) the balance goes back to zero unless there are any additional charges like wa
ter bills.

Please pay your February rent by March 1st to avoid any late fees.

Best regards,
CligProperty Team"
```

Figure 6.6: Prompting in Ollama

- **Microsoft Excel:** I used Excel to store all the scraped data, merge files, perform quick checks, and do manual validation of the datasets.

- **Microsoft Word:** I used Word to write Research and Document reports, document the 10-month project timeline, prepare validation reports, and organize all my findings. These tools and technologies helped me work faster, keep data clean and accurate, and deliver high-quality results for the CliqProperty project.

6.4 User Guideline

To maintain the functionality and accuracy of the software and data pipelines developed, users should adhere to the following instructions:

- **Extraction Protocol:** When using the school data scraper, always query in bunches of 10 school IDs to prevent site blocking and ensure stable HTML/JSON payloads.
- **Validation Execution:** Immediately after loading or concatenating new data (old-df and result-df), the user must run uniqueness checks to ensure no ID collisions and verify that all percentage-based columns remain within the 0–100 range.
- **Lead Management:** For the Australian lead database, users should filter agents by State → Suburb and prioritize "Principal" or "Director" roles for initial CRM outbound campaigns.
- **AI Copilot Testing:** Before production deployment, the user must stress-test the refusal responses using adversarial prompts to ensure the AI does not bypass established approval doctrines or attempt "money moves".
- **Compliance:** Ensure all crawling activities follow polite crawling standards, including internal rate limiting and respecting the robots.txt guidance of target directories.

6.5 Impact of the Internship

My internship experience at CliqPack Ltd. was highly beneficial both in terms of skill acquisition and direct contribution to the company's operational

goals. Web scraping is a robust method for collecting and analyzing large amounts of data, providing firms with a clear speed-to-market edge over their competition. [Thomas & Mathur \(2019\)](#) I was actively involved in building data pipelines and supporting the development of the CliqProperty AI Copilot.

By designing automated web scraping systems and validating large datasets for over 2,284 Australian schools, I helped create clean and reliable data assets. This directly supported CliqPack's mission of delivering accurate, data-driven solutions for real estate and education sectors, enabling faster decision-making and smoother Australian market expansion. My work on the ETL pipelines, polite crawling techniques, and lead database design improved the company's ability to collect and manage high-quality data ethically and efficiently. Additionally, the strategic Research and Development. I conducted — including public dataset layering and the 10-month execution timeline for the AI Copilot provided a strong foundation for the upcoming LLM-powered advisory system. On a personal level, this internship significantly strengthened my skills in web scraping, data engineering, AI research, and project planning. Working with real industry tools such as VS Code, ChromeDriver, Ollama, and Pandas boosted my confidence and gave me practical experience that goes far beyond classroom learning. It has also given me a clear direction for my future career in data engineering and AI-powered PropTech solutions. Overall, the internship has been a valuable bridge between my academic studies and professional life, preparing me well for future opportunities in the technology industry.

Chapter 7

Conclusions and Recommendations

7.1 Conclusion

The internship at Cliqpack Limited has been a transformative experience, effectively bridging the gap between my academic studies at the University of Liberal Arts Bangladesh and the professional demands of data science engineering. During this tenure, I successfully delivered high-impact data assets, including a validated repository for 2,284 Australian schools that covers detailed ICSEA values, staff statistics, and financial recurrent funding. Furthermore, my contributions to the “Cliqproperty” AI Co-pilot project—through sourcing over 8,000 property listings and establishing a 10-month execution roadmap have provided a strategic foundation for the company’s expansion into PropTech. The successful deployment of the VIC pilot dataset for Australian real estate agencies demonstrates the practical utility of the automated pipelines I constructed. Ultimately, this internship allowed me to master the end-to-end data lifecycle, from precise web extraction to “ledger-safe” AI advisory planning.

7.2 Limitations

Despite the successful completion of the assigned tasks, several limitations were encountered that impacted the workflow and system performance:

- **Extraction Constraints:** Due to site-level protections on the MySchool directory, data could not be scraped in bulk and required processing in small batches of 10 school IDs, which extended the project timeline. BeautifulSoup cannot extract all the data available to not compromise on speed. [Abodayeh et al. \(2023\)](#)
- **Hardware Performance:** Local LLM benchmarking for the Llama 3.1 8B model using Ollama revealed significant latency when processing large prompts on CPU-first architecture.
- **Data Completeness:** A primary constraint was the unavailability of agent email addresses directly from public directories, necessitating secondary enrichment strategies that were outside the initial scraping scope. Extraction of hidden web data is a major challenge nowadays because of the autonomous and heterogeneous nature of hidden web content. [Thomas & Mathur \(2019\)](#)
- **AI Safety Boundaries:** To maintain a "ledger-safe" environment, the AI Copilot had to be strictly limited by a refusal taxonomy, preventing it from performing any "money moves" or "ledger writes," which inherently limits its operational autonomy in financial contexts.
- **Site Dependency:** The reliance on XPath and CSS selectors makes the data pipelines sensitive to any changes in the HTML structure of the source websites.

7.3 Recommendations

Based on my experiences and the challenges faced during the internship, I propose the following recommendations for Cliqpack Limited:

- Implement automated monitoring tools for the scraping pipelines to keep datasets fresh and reduce manual effort.
- Provide formal training to new interns at the beginning of their placement on company tools, ethical scraping practices, and data validation standards.

- Increase intern involvement in requirement analysis and project planning sessions to gain more practical exposure and fresh ideas.
- Maintain proper documentation and records of all previous scraping scripts, validation rules, and timelines to ensure consistency across future projects.
- Explore safe and ethical data enrichment methods while strictly following privacy regulations to make the datasets even more valuable

Chapter 8

Acknowledgments

I would sincerely like to say that both my academic and industry supervisors and the team members at the organization have been supportive. They are guiding me and encouraging me to be my constant support and guide during my internship. Their insights and constructive feedback were very valuable to my learning and development as a professional. I also owe my institution the privilege of being able to have a hands-on experience in a real-world setting.

I would particularly like to acknowledge my parents, who were always supporting, motivating, and encouraging me, which allowed me to complete this internship. Their faith in me has been my strength. Without all this help and collaboration, this would not have been accomplished.

Appendix



Figure 8.1: Appointment Letter

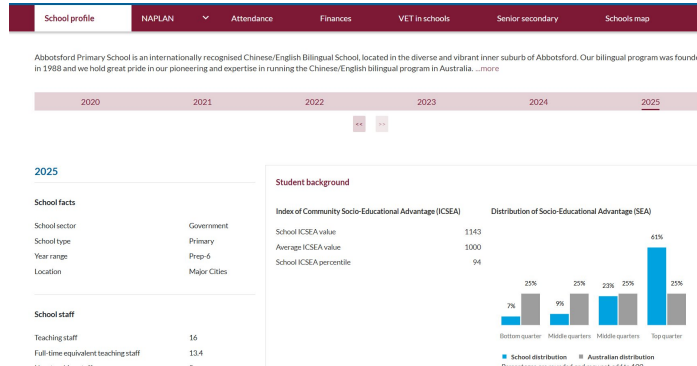


Figure 8.2: School Profile Website

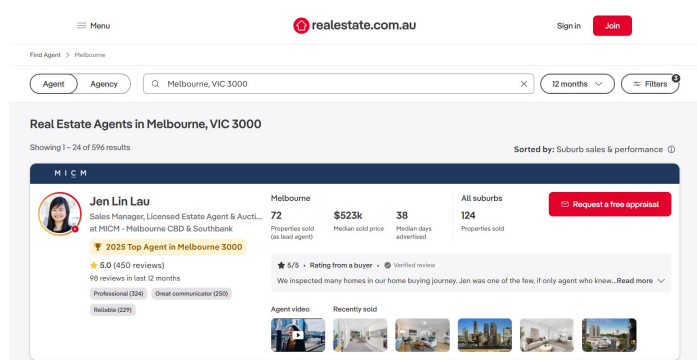


Figure 8.3: Real Estate Website of Australia

References

- Abodayeh, A., Hejazi, R., Najjar, W., Shihadeh, L., & Latif, R. (2023). Web scraping for data analytics: A beautifulsoup implementation. In *2023 sixth international conference of women in data science at prince sultan university (wids psu)* (pp. 65–69).
- Australian Curriculum, Assessment and Reporting Authority (ACARA). (2026). *My school*. <https://www.myschool.edu.au/>. (Accessed: 2026-04-21)
- Chapagain, A. (2023). *Hands-on web scraping with python: Extract quality data from the web using effective python techniques*. Packt Publishing Ltd.
- Khder, M. A. (2021). Web scraping or web crawling: State of art, techniques, approaches and application. *International Journal of Advances in Soft Computing & Its Applications*, 13(3).
- REA Group. (2026). *realestate.com.au*. <https://www.realestate.com.au/>. (Accessed: 2026-04-21)
- Thomas, D. M., & Mathur, S. (2019). Data analysis by web scraping using python. In *2019 3rd international conference on electronics, communication and aerospace technology (iceca)* (pp. 450–454).
- vanden Broucke, S., & Baesens, B. (2017). *Web scraping for data science with python*. Obtenido de KDnuggets: <https://www.kdnuggets.com/2017/12/baesens-web>
- Wahed, M. A., Alzboon, M. S., Alqaraleh, M., Ayman, J., Al-Batah, M., & Bader, A. F. (2024). Automating web data collection: Challenges, solutions, and python-based strategies for effective web scraping. In *2024 7th international conference on internet applications, protocols, and services (netapps)* (pp. 1–6).