Web Scraping Using IOT-Based Raspberry PI

Bhavya Shah
Department of Computer Engineering and Techonlogy
Dr. Vishwanath karad MIT World Peace University
Pune, Maharastra
bhavyashah16@outlook.com

Lakshya Upadhyaya
Department of Computer Engineering and Techonlogy
Dr. Vishwanath karad MIT World Peace University
Pune, Maharastra
lakshyau15@gmail.com

Gautam Sharma
Department of Computer Engineering and Techonlogy
Dr. Vishwanath karad MIT World Peace University
Pune, Maharastra
sharma.gautam0905@gmail.com

Samarth Patel
Department of Computer Engineering and Techonlogy
Dr. Vishwanath karad MIT World Peace University
Pune, Maharastra
samarthpatel23104@gmail.com

Dr. Vitthal Gutte
Department of Computer Engineering and Techonlogy
Dr. Vishwanath karad MIT World Peace University
Pune, Maharastra
vitthalgutte2014@gmail.com

***Abstract*—The proliferation of machine learning (ML) models has further driven the demand of massive and high-quality datasets, yet alternative datasets collection approaches may be cost- and resource-intensive. The current paper presents a novel, low-power Internet of Things (IoT) system to automatically scrape websites dynamically to generate ML-ready datasets, based on a limited resource edge-device. It is implemented on a Raspberry Pi edge node which gathers information automatically on dynamic websites using python packages Selenium and BeautifulSoup. This piece of work also features an active health monitoring device to facilitate resilience on prolonged working hours. The temperature of the CPU is being continuously measured and relayed over the lightweight MQTT protocol to the ThingSpeak cloud to be visualized and analyzed in real time and proves the effectiveness of passive thermal control provided by a heat sink. This two-fold design is an example of successful data-gathering of structured data types that can be trained in ML pipelines, at the same time the collection equipment is not damaged. The proposed framework confirms the possibility of low-cost edge devices to support more advanced data acquisition, and demonstrates a scalable, low-energy consuming, and allowable provisioning of researchers and developers to perform the development of custom datasets.**

***Index*—Internet of Things (IoT), Edge Computing, Web Scraping, Raspberry Pi, Dataset Creation, Machine Learning (ML), Thermal Management, MQTT**

I. INTRODUCTION

Artificial Intelligence (AI) and Machine Learning (ML) have been widely used in a sudden manner that has mostly defined the phenomenon of the fourth industrial revolution. The models largely depend on the quantity and quality of the data used in training to be successful. This has led to the capacity to acquire succinct large domain-specific datasets becoming a problem that is challenging and has to be solved. Scraping data, especially at the web scale, can be highly expensive in terms of computing power and cost, which can impede researchers and students to join the discussion in more traditional ways of scraping data.

This paper acknowledges the need to have an easier and more affordable means of not data acquisition in the broader framework of 'IoT' and helps! The technology of edge computing may meet those requirements. This paper presented a low power network that used a Raspberry Pi as a fully independent edge node that performs web scraping and that can be accessed via the internet. The system will be capable of extracting structured information on the web and capturing the data with the aid of Python and packages Selenium and BeautifulSoup that can be utilized in an ML pipeline.

However, the use of intensive apps in devices with limited resources is posing an essential operating issue to the limelight — thermal management. The constant operation of applications programs may lead to higher temperature of the CPU and lower performance and degradation of the hardware. In order to deal with this challenge, our framework employs a real-time health monitoring element. The temperature of the device is transferred by the MQTT protocol to the IoT cloud platform ThingSpeak, which makes it remotely available and gives the opportunity to confirm the influence of using an integrated passive heat sink.

The paper had the following significant contributions:
1) the architecture of an IoT system with dual use, which will enable data collection and device maintenance;
2) the proof of the creation of ML datasets using a low-cost edge device;
3) integration of a MQTT based monitoring system to facilitate operational stability.

II. Literature Review

Overview:
The high pace of Internet of Things (IoT) evolution has boosted the research of lightweight communication, edge-computing system, and system-level management. The most common protocols such as MQTT, CoAP, DPWS, and XMPP are also briefly discussed including the explanation of QoS capabilities of MQTT, small header size, and publish/subscribe pattern, and absence of discovery of MQTT is explained[3]. These security analyses consist of authentication, authorization based on ACLs, the payload encryption, tradeoff of TLS/DTLS, and the link-layer encryption (LLSec) in which hardware acceleration is applied [3].

Raspberry pi as an Edge device:
Edge computing is essential for future IoT (ECDriven- IoT) systems to achieve low latency, low energy consumption, and high scalability, addressing the flaws of remote cloud computing[8].

Raspberry boards are presented as low-power, low-cost SBCs that could be utilized in edge processing to reduce bandwidth bandwidth of the cloud-based design [4]. Pi-based image processing, service deployment, PaaS clusters, and real-time lightning are several works that are taken into consideration. Real time measurements of a Pi3B+ over time show that there is constant CPU load (around 25 per cent to 30 per cent) and slight rise in temperature during real time signal processing, which can also be regarded as viability to flash-flood forecasting. Comparative studies of the different Pi models indicate that Pi 4B can do more than Zero -W can do little.

Raspberry Pi serves as the central computing unit responsible for gathering, processing, and securely transmitting essential sign data (like heart rate and eye blink frequency)[11]. The raspberry pi functions as a Data Acquisition Console (DAC), interfacing with sensor suits and utilizing its inbuilt LAN port for connectivity[7].

High-performance processors:
Thermal control:
Thermal Assist Unit (TAU) is an on-chip micro sensing, dual-threshold registration, and interrupt controller that offers an ability to control the temperature fine-grainedly [5]. Together with the instruction cache throttling and sleep modes TAU has an optimum 1 2 derating of 7 degree C temperature at its peak performance, better than naive clock-scaling. Results of silicon result show process corner correction of -1 degree C and power variance.

IoT-Based Digital Signage:
To deliver remote, context-sensitive advertisements and real-time information (traffic, weather), the signage system is driven by a Raspberry instead of microcontroller boards, a Node.js server, web APIs, and optional DHT11 sensors [2]. The option of remote control, by a web control panel, enables to plan and update the content in real time.

Application Domains (Agriculture, Transportation, Healthcare):
Pi-enabled IoT are deployed in surveyed literature to introduce such applications as intelligent transportation (driver aid, parking, road-condition monitoring, fleet tracking) [1], smart agriculture (greenhouse climate control, irrigation, plant fertilizer, etc.), and health care (vital-sign monitoring, non-invasive glucose monitoring, elderly support). The protocols are lightweight and have edge compute functionality and powerful thermal or security enforcement across these regions, which make up a single research environment of scalable IoT solutions.

IoT Data Platform, Visualization, and Alert Systems:
ThingSpeak is a popular IoT Platform-as-a-Service (PaaS) that enables users to easily create channels to receive, host, and visualize sensor data[6]. ThingSpeak is utilized for securing storing, visualizing, and managing real-time data from wearable sensors for continuous monitoring[11]. The platform enables healthcare practitioners to access data remotely and generates warnings and alerts based on the pre defined criteria[11].

ThingSpeak is used as a smart community monitoring platform, which works on the MQTT as the data transmission channel[9]. The integrated apps of the platforms (including React, ThingHTTP, TalkBack, TimeControl, and TweetControl) play a significant role in issuing a warning (e.g. posting a fire warning on Twitter) and providing remote control of devices through time or Twitter messages[9].

Lightweight and Secure Communication in IoT:
MQTT (Message Queue Telemetry Transport) has been found to be a very lightweight publish messaging transport, which is IIoT-friendly and has been designed to work well with high latency networks or unreliable networks[3]. The MQTT protocol is suggested when a smooth and fast flow of patient data should be achieved[3]. Among the frequently used IoT communication protocols are MQTT, which is oriented more towards remote access and uses a minimal amount of power[1].

III. System Architecture & Methodology

A. Hardware Architecture:

The hardware design of our proposed framework is a stand-alone and self-contained platform, with Raspberry Pi single-board computer at its centre. The entire hardware setup is

shown in Fig. 1.

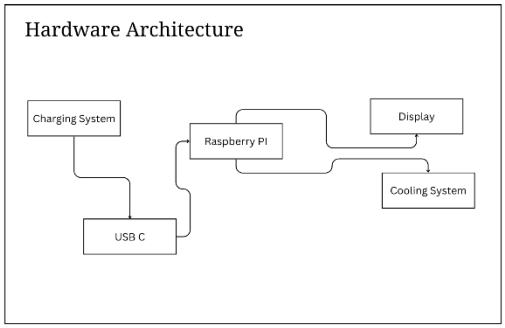

Fig 1. Hardware Architecture

This framework hardware architecture has been designed as a strong self-sufficient system, the heart of which is the Raspberry Pi. The given single-board computer was chosen to act as the central processing unit because of their appropriate combination of computational power, low power consumption and wide General-Purpose Input/Output (GPIO) interface availability, that are necessary to interface with the specific application-oriented hardware.

To support extensive scraping of data over long durations, which is an essential quality of data scraping tasks, the system is supplied by a highly stable Charging System through a modern USB-C connection. This design is not accidental because it eliminates the chances of system breakdown or corruption of data due to power variations, thus, maintaining the soundness of the data collection process.

The Raspberry Pi also manages two main peripherals that help to increase the functionality and reliability of the system greatly. A user interface, such as a dedicated Display, is essential not just during the first configuration of the system and real-time monitoring of its status, but also in on-the-fly debugging, which does not require having a second computer with which the device can communicate via SSH. Most importantly, dynamic thermal control is ensured by a dynamically controlled Cooling System, like a variable-speed fan, that is actively controlled by its GPIO. In contrast to a passive heat sink, which is a static and passive system, this active system can have intelligent response to computational load. The Pi is also programmable to turn the fan on only when certain temperature limits have been surpassed and therefore CPU throttling will not occur and optimum performance will be guaranteed during heavy scraping.

Together, these has chosen elements create a small and complete hardware system, able to serve as a self-sufficient and flexible tool of advanced data collection.

B. Software Architecture:

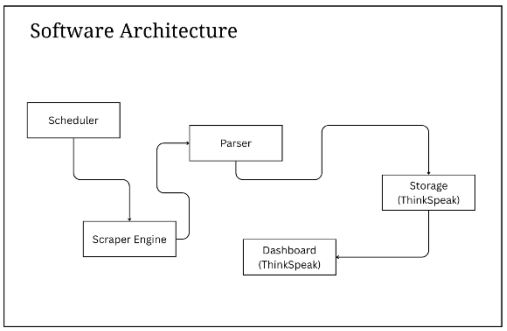

Fig 2. Software Architecture

The software is fully automated to automatically extract data on websites as per clear steps. Everything begins with a Scheduler that is similar to an alarm clock and it informs the system to begin scraping. This enables it to operate on its own. The Scraper Engine springs into action when it is time to get away. It employs two primary tools, Selenium to manipulate a web browser and navigate through the websites as a human person would, and BeautifulSoup to extract the code of the website and extract the raw information.

After collecting the raw information, the Parser is used to clean the information. It works as a filter and retrieves only the detailed information that we want, such as game titles or prices, and sorts them accordingly. The second step is the transmission of this clean data to ThingSpeak. ThingSpeak does two significant things: it archives the data to enable us to see it over time, and it displays the information on real-time Dashboard of charts and graphs, and we can easily see and comprehend the outcomes at a glance.

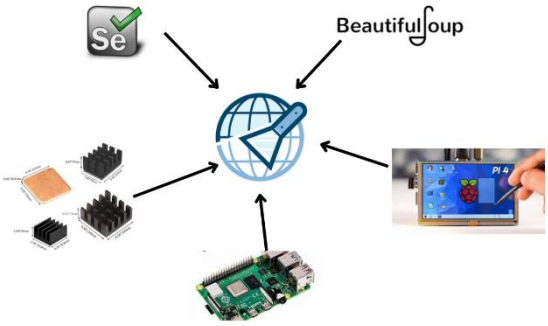

Fig 3. Graphical Abstract

The technological stack for this project integrates specific hardware and software to create an efficient, independent web scrapping system. The main part of the hardware is Raspberry Pi, a low cost single-board computer that serves as the main processing unit. To assure stable performance during heavy

tasks, the system includes heat sinks for passive thermal management. The primary user interface for interacting with the collected data is the cloud based ThingSpeak dashboard. On the software side, the data collecting is powered by Selenium, which automates a web browser to handle dynamic websites (like JavaScript related), and BeautifulSoup which parses the HTML to collect the required information. This setup allows Raspberry Pi to work as a headless data collection node, while users can analyze the results and visualize in real time through ThingSpeak interface.

C. Data Acquisition Module:

This module works on the removal of information on the target site. It is written in Python 3 and utilizes two important libraries:

1. Selenium: It is used to automate a web browser (Chromium). This is necessary in the management of modern and dynamic websites that use JavaScript to load pages. The script loads the target URL, executes activities such as scrolling or clicking, and gets the fully loaded page source.
2. BeautifulSoup: When the page source is retrieved with the help of Selenium, BeautifulSoup is applied to the HTML document to extract the HTML. It offers an effective means of traversing the parse tree to pick out particular data points by tagging, classes, and ID.

The obtained data (e.g., product names, prices, reviews) is processed, cleaned, and arranged into a Pandas DataFrame, which is further stored in the local disk as a CSV file. This format can be used across all types of machine learning systems such as Scikit-learn and TensorFlow.

D. Health Monitoring Module:

1. This will be a parallel module that operates as a different process or thread to the scraping task. Its only role is to provide the stability of the Raspberry Pi.
2. Temperature Sensing: This is a Python script that reads the internal temperature of the CPU of the Raspberry Pi periodically out of the system files under /sys/class/thermal/thermal_zone0/temp.
3. MQTT Transmission: The temperature value is delivered to a particular subject on a public or private MQTT server. The lightweight Paho-MQTT client library is used to do this. Publication is done after a predetermined time (e.g. after every 60 seconds) in order to prevent congestion on the network.
4. Cloud Visualization The ThingSpeak platform will be set up to subscribe to the MQTT topic. ThingSpeak is an IoT analytics service where it is possible to record and visualize data streams in real-time. An open channel is created and a field of temperature is put, which then automatically graphs the incoming data during the time

IV. Implementation & Results

The framework was implemented and tested to validate its two primary functions. The target for the web scraping task was a collecting Game details from gaming website and another for Football website to collect data related to matches and scores between any two teams.

A. Data Acquisition Results:

The scraping program was run over a time period of one hour. The system was able to navigate and identify the data required fields in several pages and put them together in a structured CSV file. A very small sample of the resulting dataset is presented in Table I, and it indicates that it is well-organized and can be used in the purpose of the ML.

Title	Description	Developer(s)	Publisher(s)	Release	Genre(s)	Mode(s)
30XX	Following c	Batterystap	Batterystap	Windows \	Rogue-lite ,	Single-player , multiplayer
.hack//G.U.	This is part of the list of		Nintendo Switch	games.		
Achilles: Le	Players con	Dark Point (	Dark Point (	Windows,Pl	Action role-	Single-player
Ad Infinitun	Players con	Hekate	Nacon	14-Sep-23	Survival hor	Single-player
Advance W	Advance W	WayForwar	Nintendo	21-Apr-23	Turn-based	Single-player , multiplayer
AEW Fight F	AEW Fight F	Yuke's	THQ Nordic	29-Jun-23	Sports	Single-player , multiplayer
AFL 23	AFL 23is a 2	Big Ant Stu	Nacon	Windows, P	Sports	Single-player , multiplayer
After Us	Players con	Piccolo Stu	Private Divi	WW : May 2	Action-adve	Single-player
Against the	The game is	Eremite Ga	Hooded Ho	Windows \	City-buildin	Single-player
Age of Won	At the end (	Triumph St	Paradox Int	02-May-23	4X turn-bas	Single-player , multiplayer
Aimlabs	Aimlabs, for	State Space	State Space	16-Jun-23	Shooter	Single-player
Akka Arrh	It was relea	Llamasoft	Atari	February 21	Action	Single-player
Night Sprin	Alan Wake 2	Remedy Ent	Epic Games	27-Oct-23	Survival hor	Single-player

Fig 4. Sample of Scraped Data

B. Dashboard Visualization Result:

In order to show a real-world implementation of the data acquisition framework in action, the system was implemented to scrape an open video game database. Its aim was to gather data on new game titles and their developers that have been released recently and create a dataset to analyze them. Once the data was gathered and organized, a primary analysis was conducted in order to define the most prolific developers in the data.

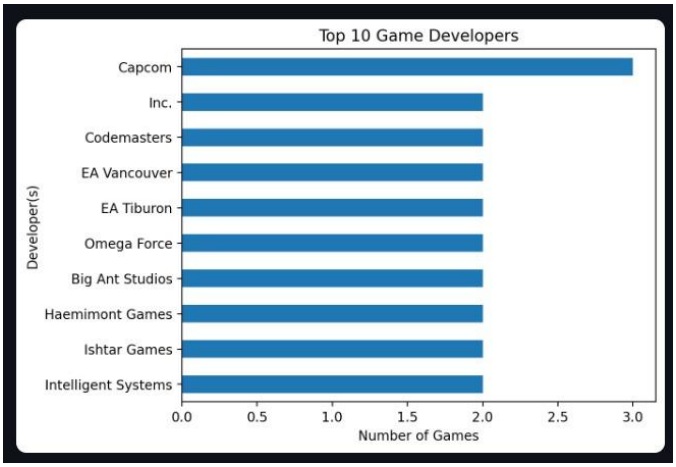

Fig 5. Bar Chart of Top 10 Game Developers from Scraped Dataset

V. Conclusion

To sum up, the current paper was able to design, implement, and validate a dual-purpose, low-power IoT framework overcoming the major problem of resource-intensive data acquisition to machine learning. Our work shows that complex, automated web scraping and data structuring can be completed successfully using a low-cost, limited-capacity edge computing node based on a Raspberry Pi. The prevalence of a real-time health monitoring feature was the key to the reliability of the system when performing long-duration tasks. A system based on the lightweight MQTT protocol used in this system offered necessary control of thermal performance of the device, and was able to avoid CPU throttling, maintaining operational stability. The symbiotic relationship between the main data collection task and the secondary self-monitoring function is the main contribution of our architecture proposal.

This study has implications worth mentioning as it offered a scalable and accessible blueprint that successfully democratizes an important phase of the machine learning lifecycle. This solution will allow researchers, students, and hobbyists to generate the custom datasets to implement innovative projects because the barriers to entry are reduced in both finance and computation. Although the implementation at present shows that the concept works on a single node, the future would incorporate the expansion of this framework to a distributed network of Raspberry Pis in order to further speed up the process of data collection. We will also improve health monitoring module to have other important measurements like memory and CPU usage so as to come up with a much stronger and smarter data collection tool.

References

[1] Hosny, Khalid M., et al. "Internet of things applications using Raspberry-Pi: a survey." International Journal of Electrical & Computer Engineering (2088-8708) 13.1 (2023).

[2] Alase, Swapnil, and Vaibhavi Chinchur. "IoT based digital signage board using Raspberry Pi 3." International Research Journal of Engineering and Technology 4.05 (2017): 310-313.

[3] S. Katsikeas et al., "Lightweight & secure industrial IoT communications via the MQ telemetry transport protocol," 2017 IEEE Symposium on Computers and Communications (ISCC), Heraklion, Greece, 2017, pp. 1193-1200, doi: 10.1109/ISCC.2017.8024687. keywords: {Protocols; Encryption; Wireless sensor networks; Payloads; Communication system security; Wireless communication; Industrial Internet of Things; IIoT; MQTT; information security; confidentiality; authentication; secure communication; Wireless Sensor Networks; WSN; Industrial Networks},

[4] Hassan, Aslinda, et al. "Performance evaluation of raspberry pi as an iot edge signal processing device for a real-time flash flood forecasting system." International Journal of Advanced Computer Science and Applications 13.10 (2022).

[5] H. Sanchez et al., "Thermal management system for high performance PowerPC/sup TM/ microprocessors," Proceedings IEEE COMPCON 97. Digest of Papers, San Jose, CA, USA, 1997, pp. 325-330, doi: 10.1109/CMPCON.1997.584744. keywords: {Thermal management; Energy management; Power system management; Portable computers; Process design; Power dissipation; Costs; Microprocessors; Thermal sensors; Logic},

[6] Thomas Zachariah, Noah Klugman, and Prabal Dutta. 2023. ThingSpeak in the Wild: Exploring 38K Visualizations of IoT Data. In Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems (SenSys '22). Association for Computing Machinery, New York, NY, USA, 1035–1040. https://doi.org/10.1145/3560905.3567766 M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

[7] Thapliyal, Aditya, and C. Kumar. "Development of data acquisition console and web server using Raspberry Pi for marine platforms." International Journal of Information Technology and Computer Science 8 (2016): 46-53.

[8] Linghe Kong, Jinlin Tan, Junqin Huang, Guihai Chen, Shuaitian Wang, Xi Jin, Peng Zeng, Muhammad Khan, and Sajal K. Das. 2022. Edge-computing-driven Internet of Things: A Survey. ACM Comput. Surv. 55, 8, Article 174 (August 2023), 41 pages. https://doi.org/10.1145/3555308

[9] Nettikadan, David, and Subodh Raj. "Smart community monitoring system using Thingspeak IoT platform." International Journal of Applied Engineering Research 13.17 (2018): 13402-13408.

[10] K. M P and D. N R, "Crop Prediction Based on Influencing Parameters for Different States in India- The Data Mining Approach," 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 2021, pp. 1785-1791, doi: 10.1109/ICICCS51141.2021.9432247. keywords: {Temperature dependence; Temperature; Vegetation; Soil; Prediction algorithms; Agriculture; Classification algorithms; crop; climate; pesticides; fertilizers; groundwater; soil; Agriculture}.

[11] Sivabalaselvamani, D. & Kempaiyan, Nanthini & Nagaraj, Bharath Kumar & Kannan, K & Hariharan, K & Mallingeshwaran, M. (2024). Healthcare Monitoring and Analysis Using ThingSpeak IoT Platform: Capturing and Analyzing Sensor Data for Enhanced Patient Care Healthcare Monitoring and Analysis Using ThingSpeak IoT Platform. 26.