IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 12, Issue 5, May 2024

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.379

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 5, May 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1205310 |

Cloud Based Web Scraping for Big Data Applications

Mr. R. Karthikeyan, Mr. S. Manikandan

Assistant Professor, Department of Master of Computer Application, Gnanamani College of Technology, Namakkal Tamil Nadu, India

PG Scholar, Department of Master of Computer Application, Gnanamani College of Technology, Namakkal, Tamil Nadu, India

ABSTRACT: Web scraping, also known as web extraction or harvesting, is a technique to extract data from the World Wide Web (WWW) and save it to a file system or database for later retrieval or analysis. As scraping is one of the major sources for extraction of unstructured data from the Internet, we have analyzed the scraping process when introduced to a bulk of data extraction.. We Proposes a cloud-based web scraping architecture able to handle storage and computing resources with elasticity on demand using Amazon Web Services(Elastic Compute Cloud and Dynamo DB). Web scraping is a technique used to extract data from websites. With the exponential growth of data on the web, web scraping has become increasingly important for collecting and analyzing large amounts of data. Traditional web scraping methods are often limited by the hardware resources available on a single machine, making it challenging to scale up for big data applications. In this project, we propose a cloud-based web scraping framework for big data applications. Our framework leverages the scalability and flexibility of cloud computing to overcome the limitations of traditional web scraping methods. By deploying web scraping tasks to a cloud environment, our framework can easily scale up to handle large volumes of data and can be dynamically adjusted to accommodate varying workloads. We implement our framework using Python and the BeautifulSoup library for web scraping, and we deploy it on the Google Cloud Platform (GCP) using Google Cloud Functions for serverless computing. We demonstrate the effectiveness of our framework by conducting experiments on a large dataset of web pages, showing that it can efficiently scrape data from thousands of web pages in parallel. The web scraping system is a powerful tool for big data applications. It enables organizations to gather, process, and analyze large volumes of web data in a scalable, efficient, and timely manner, opening up new possibilities for data-driven decision-making and insights.

KEYWORDS: Cloud Computing, Web Scraping, cloud-based web scrapper, Selenium, XPath.

I. INTRODUCTION

WEB SCRAPING

The proposed web scraping system a powerful tool for big data applications. It enables organizations to gather, process, and analyze large volumes of web data in a scalable, efficient, and timely manner, opening up new possibilities for data-driven decision-making and insights.

Today, cloud computing has become ubiquitous, with major technology companies such as Google, Microsoft, and IBM offering a wide range of cloud services. The development of cloud computing has been driven by factors such as the increasing demand for computing resources, the need for greater flexibility and scalability, and the growing importance of data analytics and big data.

This versatility grants major development was the widespread adoption of virtualization technology, which enables the creation of virtual instances of hardware resources such as servers, storage, and networking. Virtualization allows for greater flexibility and efficiency in resource allocation, making it a key enabler of cloud computing.

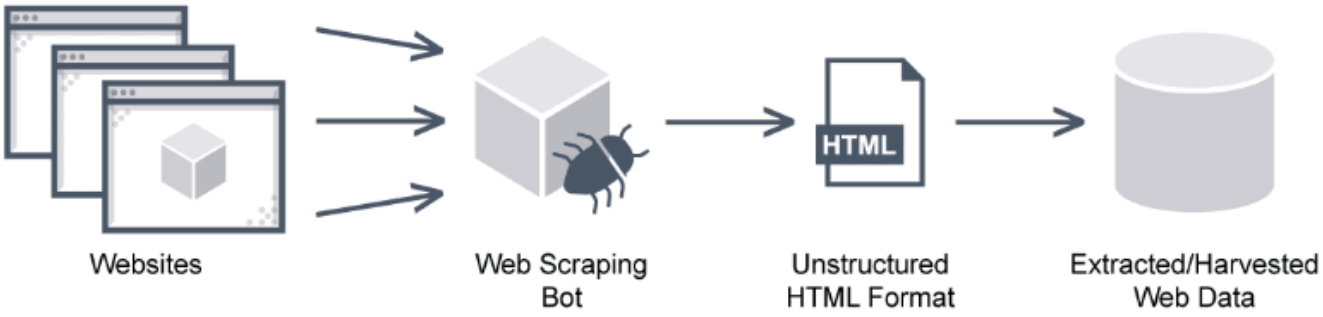
International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 5, May 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1205310 |

TECHNIQUES OF WEB SCRAPING



WEB SCRAPING SOFTWARE

Automated tools designed to extract data from websites without manual intervention. These tools typically allow users to specify the data to be extracted using a visual interface or scripting language, and then automate the process of navigating the website and extracting the data..

XPATH

A query language for selecting nodes in an XML document. XPath is commonly used in web scraping to navigate the hierarchical structure of HTML documents and extract specific elements based on their position in the document tree.

CSS SELECTORS

A method for selecting elements in an HTML document based on their CSS attributes. CSS selectors are commonly used in conjunction with web scraping libraries like BeautifulSoup to extract specific elements from web pages based on their styling.

WEB SCRAPING LIBRARIES

Software libraries that provide tools and functions for web scraping. These libraries simplify the process of web scraping by providing abstractions for common tasks like parsing HTML, navigating web pages, and extracting data, making it easier for developers to write web scraping code.

WEB SCRAPING BOT

A web scraping bot, also known as a web crawler or spider, is a software program designed to automatically navigate the internet and extract information from websites. These bots simulate human browsing behavior to access web pages, follow links, and extract data based on predefined rules or patterns. Web scraping bots are used for various purposes, including data collection, search engine indexing, price monitoring, content aggregation, market research, and monitoring website changes.

UNSTRUCTURED HTML FORMAT

To handle unstructured HTML, web scraping bots often use advanced parsing techniques and algorithms. For example, they might employ machine learning algorithms to identify and extract relevant data from unstructured text. Alternatively, they might use heuristics to analyze the layout and content of a web page to infer the structure and extract data accordingly. By leveraging advanced parsing techniques and algorithms, these bots can navigate complex web pages and extract data with a high degree of accuracy.

EXTRACTED/HARVESTED WEB DATA

Extracted or harvested web data refers to the information collected by web scraping bots from various websites. This data can be in various formats, such as text, images, or structured data like tables. Once the web scraping bot collects the data, it is typically stored in a structured format.

II. LITERATURE SURVEY

Title : Data mining on parallel database systems
Author : Mauro sousa marta mattoso nelson ebecken

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 5, May 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1205310 |

Progressing years have shown the need of a robotized cycle to find in-teresting and secret models in genuine informational indexes, dealing with colossal volumes of data. This sort of cycle proposes a lot of com-putational power, memory and plate I/O, which should be given by equivalent com-puters. Our work contributes with a solu-tion that organizes a computer based intelligence algo-rithm, parallelism and a solidly coupled usage of a DBMS system, keeping an eye on execution issues with equivalent taking care of and data crack.

Title : **Ant colony system for graph coloring problem**
Author : Malika bessedik, rafik laib, aissa boulmerka et habiba drias

In this paper, we present a first ACO approach, specifically Bug Settlement Structure (ACS) for the graph concealing issue (GCP). We executed two methodology of ACS for the GCP; advancement system and improvement procedure. Being developed methodology, the computation iterativelly assembles feasible game plans. The time of improvement is finished by a specific significant procedure for the issue, that is: Recursive Greatest First (RLF) or DSATUR.

Title : A definition of peer-to-peer networking for the classification of peer-to-peer architectures and applications
Author : Riidiger schollmeier

The vital responsibility ojthe flag, which is moving right along delineated in coming up straightaway, is to ofer a definition for Peerto-Friend coordinating and to make the differentiations to typical assumed Client/Server-structures clear. With this definition we can classrjji as of now existing frameworks organization thoughts in the Internet either as "Pure" Conveyed, or "Cream" Shared or Client/Server plan,

Title : Review of mobile banking and its evolvingtrend in india
Author : Hamia khan

With the presence of advancement, banking industry has in like manner created. The business has been using advancement. Advancement has upheld the monetary business for straightforwardness of conveying organizations. Web has furthermore shown to clear way for different ventures driving them to introduce new item offering and has displayed to be helpful for banking industry. In the present automated age, mobile phones are the fundamental strategy for getting to the web. Extended sensibility and accessibility of PDA and the ascent of mix feature phones has incited all over web use. Banks serve clients capably using various stations and branches like Robotized Teller Machines (ATM), web banking, telephone banking, and compact banking. Versatile banking has itself created from Short Message Organization (SMS) banking; adaptable applications to got biometric applications M-Banking let clients to help banking organizations 24*7. It has pushed ahead and has turned out to be important to clients and has been useful for the monetary business moreover. Anyway there are troubles especially as for security reason which banking region need to control to advance.

Title : IP-based virtual privatenetwork implementations in future cellular networks
Author : Madhusanka liyanage, mika ylianttila, andrei gurtov

Virtual Secret Association (VPN) organizations are for the most part used in the present corporate world to securely interconnect geographically scattered private association segments through temperamental public associations. Among various VPN techniques, Web Show (IP)- based VPN organizations are overpowering a result of the ubiquitous usage of IP-based provider associations and the Internet. Over latest several numerous years, the use of cell/adaptable associations has extended enormously as a result of the fast increase of the amount of convenient allies and the evolvement of media transmission progresses. Besides, cell network-based broadband organizations can give a comparative plan of association organizations as wired Internet services. Thusly, compact broadband organizations are furthermore turning out to be notable among corporate clients. In this way, the utilization of convenient broadband organizations in corporate associations solicitations to execute different broadband organizations on top of adaptable associations, including VPN organizations. This part is revolved around recognizing critical level use cases and circumstances where IP-based VPN organizations can be done on top of cell associations. Also, the makers expect the future commitment of IP-based VPNs in past LTE cell associations

MODULES

- Scraper Module
- Cloud-based Scraper Module.
- Selenium Module
- Puppeteer Module
- PyQuery Module.

International Journal of Innovative Research in Computer and Communication Engineering

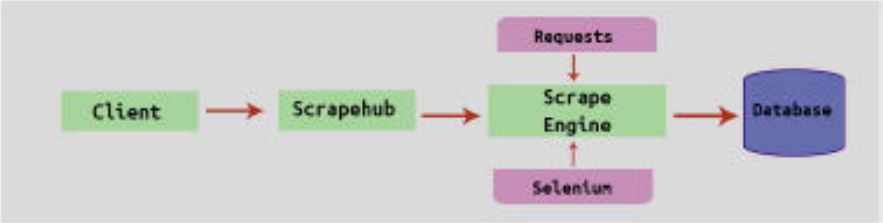
| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 5, May 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1205310 |

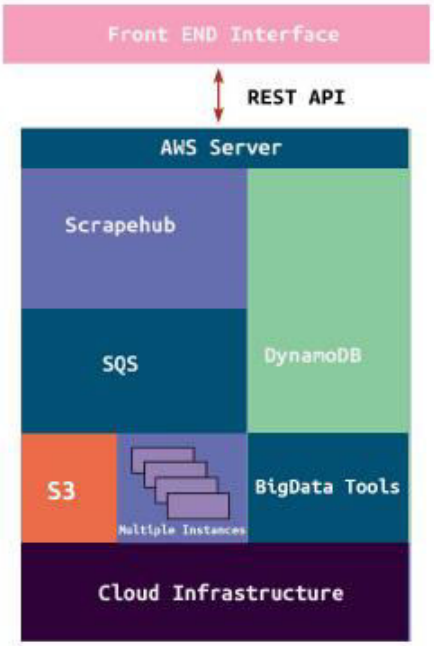
III. MODULES DESCRIPTION

SCRAPER MODULE



The scrape-hub program initiates, handles and monitors all tasks in the system. It is written in python language. Then scrape-hub starts the scraping engine, the task of which is to initiate a web page, automate it to the desired state using selenium library and parse it using request library. Different scrapers may use a different library for parsing and automate web pages.

CLOUD BASED SCRAPER MODULE



Front-end platform for the architecture will be web-browser. Front-end is connected to amazon web server in the cloud using REST API. scrape-hub, residing at cloud will handle all client requests of scraping as well as handling and monitoring events. It is also responsible for assigning the required number of instances for computing depending upon URLs chunk.

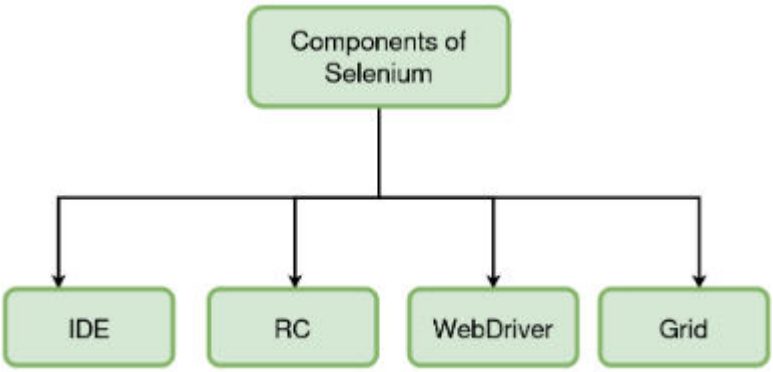
International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 5, May 2024 ||

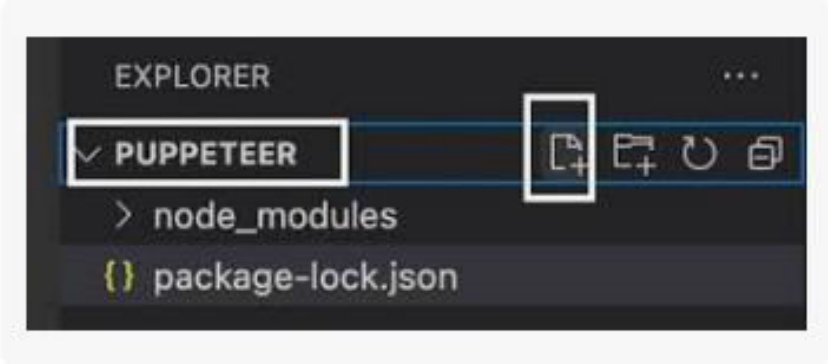
| DOI: 10.15680/IJIRCCE.2024.1205310 |

SELENIUM MODULE



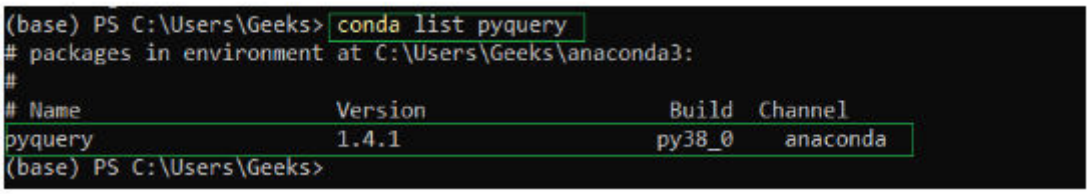
A web automation tool that can be used for web scraping. Selenium allows you to control a web browser programmatically, enabling you to interact with dynamic content and JavaScript-heavy websites.

PUPPETEER MODULE



A Node.js library for controlling a headless version of the Chrome browser. Puppeteer can be used for web scraping and provides tools for interacting with web pages and extracting data.

PYQUERY MODULE



A Python library that provides jQuery-like syntax for parsing HTML documents. PyQuery allows you to use CSS selectors to extract data from web pages easily. These modules can be used individually or in combination to build powerful web scraping tools for extracting data from websites.

IV. WEB SCRAPING CHARACTERISTICS

DATA COLLECTION

Web scraping automates the process of collecting large amounts of data from websites. It can gather data in various formats like HTML, JSON, XML, and others.

AUTOMATION

Scraping scripts or tools automate the extraction process, allowing for the collection of data at scale.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 5, May 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1205310 |

Popular tools include BeautifulSoup, Scrapy, Selenium, and Puppeteer.

PARSING

Scraping involves parsing HTML or other formats to extract relevant information. This requires an understanding of the website's structure (DOM) and possibly handling dynamic content loaded by JavaScript.

V. ALGORITHM

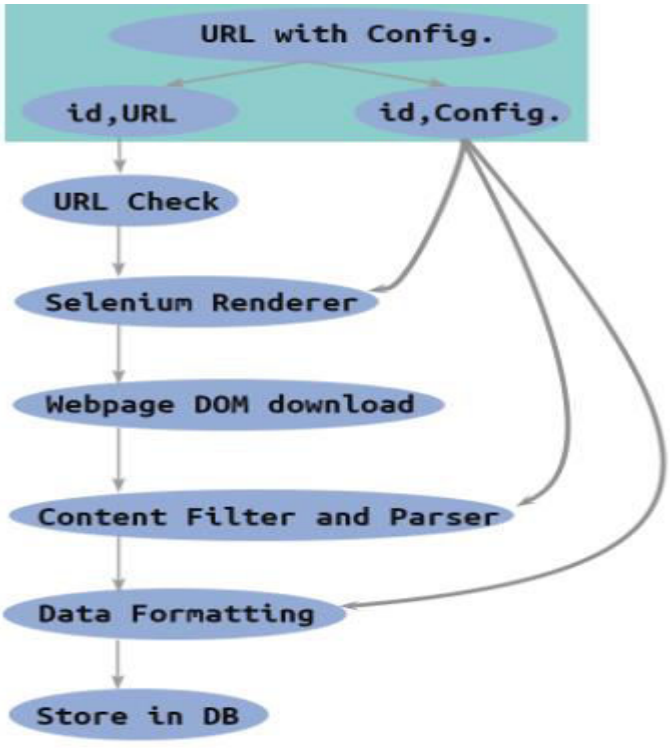
The Cloud Based Web Scraping for Big Data Applications, the algorithm is a structured sequence of steps designed to efficiently extract, process, and store large volumes of data from various websites using cloud computing resources. It begins by identifying the target websites and specifying the data to be scraped. A web scraping script is developed using tools like BeautifulSoup or Scrapy, and for dynamic content, Selenium is incorporated. The cloud environment is set up on platforms like AWS, Google Cloud, or Azure, provisioning necessary resources such as EC2 instances or Lambda functions.

VI. IMPLEMENTATION

To implement a cloud-based web scraping system for big data applications, begin by defining your objectives and identifying the target websites and specific data to be extracted. Develop the scraping script using frameworks like BeautifulSoup or Scrapy, and for handling dynamic content, incorporate Selenium. Set up your cloud environment by selecting a provider such as AWS, Google Cloud, or Azure, and create the necessary cloud resources like EC2 instances or Lambda functions. Implement the scraping logic within your script, ensuring to include error handling and retries. Distribute the scraping tasks using serverless cloud services like AWS Lambda or Google Cloud Functions for automatic scaling, or orchestrate containers with Kubernetes to manage workloads efficiently.

To bypass anti-scraping measures, use techniques like IP rotation and user-agent spoofing, and introduce delays between requests. Process and clean the extracted data in real-time using cloud data processing services like AWS Glue or Google Dataflow. Store the cleaned data in scalable cloud storage solutions such as Amazon S3 or Google Cloud Storage, ensuring it's structured for further use. Integrate this data with big data tools like Apache Spark or Hadoop running on cloud clusters for analysis.

VII. SYSTEM ARCHITECTURE



| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 5, May 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1205310 |

VIII. CONCLUSION

In conclusion, the implementation of a cloud-based web scraping system for big data applications effectively leverages the scalability, flexibility, and robustness of cloud computing to handle large-scale data extraction and processing tasks. By utilizing frameworks like BeautifulSoup, Scrapy, and Selenium, and integrating with cloud services such as AWS Lambda, Google Cloud Functions, and Kubernetes, this approach ensures efficient and automated data collection from a variety of web sources. The use of cloud-based data processing tools like AWS Glue and Google Dataflow enables real-time data cleaning and transformation, while scalable storage solutions like Amazon S3 and Google Cloud Storage facilitate seamless data management.

REFERENCES

1.R.Karthikeyan, & et all "Biometric for Mobile Security" in the international journal of Engineering Science & Computing, Volume7,Issue6, June 2017, ISSN(0):2361-3361,PP No.:13552-13555.
2.R.Karthikeyan, & et all "Data Mining on Parallel Database Systems" in the international journal of Engineering Science & Computing, Volume7,Issue7, July 2017, ISSN(0):2361-3361,PP No.:13922-13927.
3.R.Karthikeyan, & et all "Ant Colony System for Graph Coloring Problem" in the international journal of Engineering Science & Computing, Volume7,Issue7, July 2017, ISSN(0):2361-3361,PP No.:14120-14125.
4.R.Karthikeyan, & et all "Classification of Peer –To- Peer Architectures and Applications" in the international journal of Engineering Science & Computing, Volume7,Issue8, Aug 2017, ISSN(0):2361-3361,PP No.:14394-14397.
5.R.Karthikeyan, & et all "Mobile Banking Services" in the international journal of Engineering Science & Computing, Volume7,Issue7, July 2017, ISSN(0):2361-3361,PP No.:14357-14361.
6.R.Karthikeyan, & et all "Neural Networks for Shortest Path Computation and Routing in Computer Networks" in the international journal of Engineering and Techniques, Volume 3 Issue 4, Aug 2017, ISSN:2395-1303,PP No.:86-91.
7.R.Karthikeyan, & et all "An Sight into Virtual Techniques Private Networks & IP Tunneling" in the international journal of Engineering and Techniques, Volume 3 Issue 4, Aug 2017, ISSN:2395-1303,PP No.:129-133.
8.R.Karthikeyan, & et all "Routing Approaches in Mobile Ad-hoc Networks" in the International Journal of Research in Engineering Technology, Volume 2 Issue 5,Aug 2017, ISSN:2455-1341, Pg No.:1-7.
9.R.Karthikeyan, & et all "Big data Analytics Using Support Vector Machine Algorithm" in the International Journal of Innovative Research in Computer and Communication Engineering, Volume 6 Issue 9, Aug 2018, ISSN:2320 - 9798, Pg No.:7589 -7594.
10.R.Karthikeyan, & et all "Data Security of Network Communication Using Distributed Firewall in WSN " in the International Journal of Innovative Research in Computer and Communication Engineering, Volume 6 Issue 7, July 2018, ISSN:2320 - 9798, Pg No.:6733 - 6737.
11.R.Karthikeyan, & et all "An Internet of Things Using Automation Detection with Wireless Sensor Network" in the International Journal of Innovative Research in Computer and Communication Engineering, Volume 6 Issue 9, September 2018, ISSN:2320 - 9798, Pg No.:7595 – 7599.
12.R.Karthikeyan, & et all "Entrepreneurship and Modernization Mechanism in Internet of Things" in the International Journal of Innovative Research in Computer and Communication Engineering, Volume 7 Issue 2, Feb 2019, ISSN:2320 - 9798, Pg No.:887 - 892.
13.R.Karthikeyan & et all "Efficient Methodology and Applications of Dynamic Heterogeneous Grid Computing" in the International Journal of Innovative Research in Computer and Communication Engineering, Volume 7 Issue 2, Feb 2019, ISSN:2320 - 9798, Pg No.:1125 -1128.
14.R.Karthikeyan & et all"Entrepreneurship and Modernization Mechanism in Internet of Things" in the International Journal of Innovative Research in Computer and Communication Engineering, Volume 7 Issue 2, Feb 2019, ISSN:2320 - 9798, Pg No.:887– 892.
15.R.Karthikeyan & etall"Efficient Methodology for Emerging and Trending of Big Data Based Applications" in the International Journal of Innovative Research in Computer and Communication Engineering, Volume 7 Issue 2, Feb 2019, ISSN:2320 - 9798, Pg No.:1246– 1249.
16.R.Karthikeyan & et all "Importance of Green Computing In Digital World" in the International Journal of Innovative Research in Computer and Communication Engineering, Volume 8 Issue 2, Feb 2020, ISSN:2320 - 9798, Pg No.:14 – 19.
17.R.Karthikeyan & et all "Fifth Generation Wireless Technology" in the International Journal of Engineering and Technology, Volume 6 Issue 2, Feb 2020, ISSN:2395–1303.
18.R.Karthikeyan & et all "Incorporation of Edge Computing through Cloud Computing Technology" in the International Research l Journal of Engineering and Technology, Volume 7 Issue 9, Sep 2020 ,p. ISSN:2395–0056, e. ISSN:2395–0072.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 5, May 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1205310 |

19.R.Karthikeyan & et all "Zigbee Based Technology Appliance In Wireless Network" in the International Journal of Advance Research and Innovative Ideas in Education, e.ISSN:2395 - 4396, Volume:6 Issue: 5 , Sep. 2020. Pg.No: 453 – 458, Paper Id:12695.

20.R.Karthikeyan & et all "Automatic Electric Metering System Using GSM" in the International Journal of Innovative Research in Management, Engineering and Technology, ISSN: 2456 - 0448, Volume:6 Issue: 3 , Mar. 2021. Pg.No: 07 – 13.

21.R.Karthikeyan & et all "Enhanced the Digital Divide Sensors on 5D Digitization" in the International Journal of Innovative Research in Computer and Communication Engineering, e-ISSN: 2320 – 9801, p-ISSN: 2320 - 9798, Volume:9 Issue: 4 , Apr. 2021. Pg.No: 1976 – 1981.

22.R.Karthikeyan & et all "Crop Yield Prediction Based On Indian Agriculture Using Machine Learning" in the International Journal Of Engineering and Techniques, ISSN: 2395-1303, Volume:8 Issue: 4 , July. 2022. Pg.No: 11 – 22.

23.R.Karthikeyan & et all "College Bus Transport Management Web Application" in the International Journal Of Multidisciplinary Research In Science, Engineering and Technology, ISSN: 2582-7219, Volume: 6 Issue: 6, June. 2023. Pg.No: 1619 – 1625.

24.R.Karthikeyan & et all "Face Recognition Based Attendance System" in the International Journal of Innovative Research in Computer and Communication Engineering, ISSN: e 2320-9801, Volume: 11 Issue: 6, June. 2023. Pg.No: 8710 – 8717.

25.R.Karthikeyan & et all "Cloud data Deduplication System using per File parity and File Name Interpreter" in the International Journal of Advanced Research in Arts, Science, Engineering and Management, ISSN: 2395-7852, Volume: 10 Issue: 3, May. 2023.

26.R.Karthikeyan & et all "Secure Photo Sharing Social Networks Using Coverless Image Steganography Techniques" in the International Journal of Research in Science, Engineering and Technology, e - ISSN: 2319-8753, Volume: 12 Issue: 6, June. 2023.Pg.No: 8852 – 8861.

27.R.Karthikeyan & et all "Hotel Booking Mobile And Web Application" in the International Journal Of Multidisciplinary Research In Science, Engineering and Technology, ISSN: 2582-7219, Volume: 6 Issue: 6, June. 2023. Pg.No: 1823 – 1829.

28.R.Karthikeyan & et all "Campus Placements Prediction & Analysis Using Machine Learning" in the International Journal of Research in Science, Engineering and Technology, e - ISSN: 2319-8753,Volume: 12 Issue: 6, June. 2023.Pg.No: 8837 – 8841.

29.R.Karthikeyan & et all "E-Auction System Web Application" in the International Journal of Research in Science, Engineering and Technology, e - ISSN: 2319-8753, Volume: 12 Issue: 6, June. 2023. Pg.No: 8837 – 8841.

30.R.Karthikeyan & et all "Digital Gram Panchayat Services Using Cloud Based System" in the International Journal of Innovative Research in Computer and Communication Engineering, e - ISSN: 2320-9801, Volume: 12 Issue: 5, May. 2024. Pg.No: 6520 – 6527.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi crossref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details