.

**Parking Management System using OpenCV**

Submitted in partial fulfillment of the requirements for the award of degree of

**BACHELOR OF ENGINEERING**

**IN**

**COMPUTER SCIENCE & ENGINEERING**



**Submitted to:**

**Project Mentor name**

**Er. Divya K**

**Submitted By:**

**18BCS16-B**

**Priyanshu Singh**

**18BET1057**

**DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING**

**Chandigarh University, Gharuan**

**May 2022**

### CERTIFICATE

This is to certify that the work embodied in this Project Report entitled **“Parking Management System using OpenCV”** being submitted by **“Priyanshu Singh ” -** UID **“18BET1057 ”, 8**th Semester for partial fulfillment of the requirement for the degree of **“Bachelor of Engineering in Computer Science & Engineering ”** discipline in “ **Chandigarh University** ” during the academic session Jan-July 2022 is a record of bonafide piece of work, carried out by student under my supervision and guidance in the **“ Department of Computer Science & Engineering ”, Chandigarh University.**

**APPROVED & GUIDED BY:**

**Prof Divya K**

**Internship Certificate**

**Na(Not yet completed the internship)**

### DECLARATION

### I, student of Bachelor of Engineering in Computer Science & Engineering, 8th Semester , session: Jan – July 2022, Chandigarh University, hereby declare that the work presented in this Project Report entitled “Parking Management System using OpenCV ” is the outcome of my own work, is bona fide and correct to the best of my knowledge and this work has been carried out taking care of Engineering Ethics. The work presented does not infringe any patented work and has not been submitted to any other university or anywhere else for the award of any degree or any professional diploma.

**Student details and Signature**

**Priyanshu Singh(18bet1057)**

**APPROVED & GUIDED BY:**

**Prof Divya K**

**ACKNOWLEDGEMENT**

We express our deepsense of gratitude to our respected and learned guide, **Prof Divya K** for their valuable help and guidance we are thankful to her for the encouragement.

We are also thankful to all the other faculty & staff members of our department for the kind cooperation and help.

Lastly, we would like to express our deep appreciation towards our classmates and our parents for providing us the moral support and encouragement.

**Name :** Priyanshu Singh

**Table of Contents**

**Certificate ii**

**Internship Certificate Provided by Internship Organisation iii**

**Declaration iv**

**Acknowledgements v**

**Table of Contents vi**

**List of Figures vii**

**List of Tables viii**

**List of Acronyms ix**

**Terminology used x**

**Abstract xi**

**1. Introduction**

1.1. Big Data 1

1.2. Big Data Technologies 3

1.2.1. Hadoop and MapReduce 4

1.2.2. Scala 5

1.2.3. Linux 6

1.2.4. Apache Nifi 7

1.2.5. Python 7

1.2.6. HDFS 7

1.2.7. HBase 8

1.2.8. NoSQL 8

1.2.9. Sqoop 9

1.2.10. Kafka 10

**List of Figures**

|  |  |  |
| --- | --- | --- |
| **Figure No.** | **Title** | **Page No.** |
| 1. | ER diagram of object acquisition and Tracking | 58 |
| 2. | Data Flow Diagram of parking system | 59 |
| 3. | Vehicle detection by background substraction | 57 |
| 4. | Track of Car | 57 |
| 5. | Application image | 56 |

**ABSTRACT**

This paper covers image development enhancements based on parking management. This project will work based on the ideas of the Background Removal algorithm. The use of this algorithm will be used as a map tool to reduce vehicle error. Now car parking is a big problem in a smart city. Due to the increase in traffic problems, the clever parking system operated using opencv provides an easy way out. Object Discovery has had a major impact about how the world has adapted to artificial intelligence in recent years. Other popular acquisition algorithms are area-based accumulative neural networks, one-time multiple-box detectors (SSDs), and Look Together (YOLO). Among these SSDs has better accuracy, while YOLO performance is better when it comes to speed provided with the selected accuracy. In-depth learning includes SSDs and Mobile Nets for easy access and use for tracking. This algorithm enables accurate detection while not compromising performance. All app functionality is based on the acquisition of an object in a particular location such as whether a rectangular space is completed or not. When found complete it means the site is empty and look for other options.

**Chapter 1**

**INTRODUCTION**

**Company Profile**

**Company Name - Cognizant**

Cognizant Technology Solutions Corporation is a **professional services company**. The Company operates through four segments: Financial Services, Healthcare, Products and Resources, and Communications, Media and Technology. The Financial Services segment includes banking, capital markets and insurance companies. Cognizant (Nasdaq-100: CTSH) engineers modern businesses. We help our clients modernize technology, reimagine processes and transform experiences so they can stay ahead in our fast-changing world. Together, we’re improving everyday life

**Role - Programmer Analyst**

Programmer analysts work at various organizations. They design, code, and test new programs. They must document programs that they write. Documentation should include flowcharts, layouts, diagrams, charts, code comments, and revision dates. They debug, troubleshoot, and maintain source code related to various computer programs. Programmer analysts may also be required to customize computer programs that were bought from vendors to meet their organization’s specific needs. They must know computer algorithms, formal logic, and application domain. They may be required to work with vendors, in order to test and approve their products.

They must usually have a bachelor's degree in computer science, information technology, or other relevant fields. They must have previous work experience in their field. Computer programmers must have thorough knowledge of Java/Java EE, C, C++, C#, .NET, Java Script, and other programming languages. Some companies may accept candidates with significant years of experience in lieu of a bachelor's degree requirement. They must have excellent oral and written communications skills.

They must adhere to confidentiality rules related to system access, data access, and security rules. They must be able to work independently with minimal or no supervision and as part of the team. They must keep themselves up to date with various new computer technology and programming languages that are available in the marketplace. They may have to travel to vendors' or clients' offices. They may be required to attend relevant classes, workshops, or conferences. They may be required to coach new or more junior programmer analysts.

**Locations**

**Chennai**

Cognizant Technology Solutions India Pvt Ltd,  KITS Campus (CKC) , Plot No. 1 & 2, Elcot IT Park, Survey, No. 602/3, Shollinganallur Village, Tambaram Taluk, Kancheepuram District - Chennai - 600119, India

**Kochi**

Cognizant Technology Solutions, 3rd Floor, Athulya Building, Infopark SEZ, Kakkanad, Kochi - 682030

**Coimbatore**

Cognizant Technology Solutions India Pvt. Ltd.,C3, CHIL Special Economic Zone(SEZ), Keeranatham Village,, Saravanampatti via Coimbatore - 641035

**Hyderabad**

Cognizant Technology Solutions India Pvt Ltd. Plot No: 129 to 132, APHB Colony, Lords Conference Hall, 8th Floor,Phase-3(Block-1), DLF Building, Opp to Old CMC Office, Gachibowli, Hyderabad-500032

**Mumbai**

Cognizant Technology Solutions India Pvt. Ltd. , 2nd Floor, SEZ, Building No.9 , Mindspace-Airoli, Thane Belapur Road, Navi Mumbai, Thane-400708

**Gurgaon**

Cognizant Technology Solutions, BLD 3 - 1st Floor, HR Bay, Candor Infospace, Sector-48, Near Subhash Chowk, Gurgaon, Haryana - 122001

**Mysore**

Silver Spirit Tech Park, 316 SPI City, Hebbal Industrial Area,Mysuru. 570001   
OB venue – Board Room.

**Bangalore**

Cognizant Technology Solutions India Pvt. Ltd. Manyata Embassy Business Park,,Mahogany(F2)-SEZ, Hebbal Outer Ring Road,,Rachenahalli, Bangalore, 560045

**Pune**

Cognizant Technology Solutions India Pvt. Ltd.IDB Auditorium, Plot No. 26 & 27, Rajiv Gandhi Infotech Park,Hinjewadi Ph-1, MIDC, Pune - 411057

**Gurgaon**

Cognizant Technology Solutions, Tower 3, 1st Floor, Unitech Reality Projects Limited,Village Tikri, Sector 48, Near Subhash Chowk, Gurgaon, Haryana  Pincode – 122001

**Kolkata**

****Venue - 1 (Lateral)****

Cognizant Technology Solutions India Pvt. Ltd.,Plot No. # GN 34/3, Salt Lake Electronic Complex,Sector – V, Salt Lake City,Kolkata – 700091

****Venue - 2 (Campus) - Bantala****

Cognizant Technology Solutions India Pvt. Ltd.Plot- 27, Kolkata IT Park, SEZ,Mouza: Gangapur, J.L.NO.35, GP: Taradah,PS: Kolkata leather Complex,Dist: South 24 Paraganas

**Noida**

Cognizant Technology Solution India Pvt. Ltd,Upper Ground Floor, 1st , 2nd Floor, Building No-10,IT/ITES SEZ of M/s Sea View Developers Ltd, Plot No 20 & 21, Sector-135, Noida- 201301

**Mangalore**

Cognizant Technology Solutions India Pvt Ltd, 7th Floor, Manasa Towers, MG Road, Kodialbail, Mangalore - 575003

**United States**

Center Valley, Upper Saucon Township

Bentonville, AR

United States

300 SW 24th St

**Teaneck (HQ), NJ**

**United States**

500 Frank W Burr Blvd

**Charlotte, NC**

8520 Cliff Cameron Drive Suite 200 Charlotte, NC 28269

Departments: IT, Other, Engineering, Art/Design, Finance, Accounting, Corporate

**Bloomington, IL**

Departments: IT, Other, Marketing, Engineering

**Houston, TX**

and Gas 3010 Briarpark Drive, Ste 250 Houston, TX 77042

Departments: Other, IT, Finance, Engineering, Marketing

**Juno Beach, FL**

Departments: Other, IT, Finance, Art/Design, Marketing, Engineering

**Johns Creek, GA**

Departments: IT, Other, Art/Design, Engineering, Marketing

### NEW ZEALAND

Suite 2720, Level 2  
Pwc Building, 188 Quay Street  
Auckland, 1010

### HONG KONG

20/F, Golden Center  
188 Des Voeux Road Central  
Sheung Wan

### KINGDOM OF SAUDI ARABIA

Office No. 10, Tower 1  
8th Floor, Tatweer Tower  
King Fahad Road, Riyadh

### MALAYSIA

C-11-01, iTech Tower, Jalan Impact,  
Cyber 6, 63000 Cyberjaya,  
Selangor Darul Ehsan

### SINGAPORE

1 Changi Business Park Crescent  
Plaza 8@CBP, Tower A  
#07-04/05/06, Singapore - 486025

### UNITED ARAB EMIRATES

Office No. 2301,  
Al Thuraya Tower 2  
Dubai Internet City, P.O. Box 500378

### ABU DHABI

Makeen Tower, Office no. 802,  
Al Salam Street – Tourist Club Area,  
Al Zahiya, Abu Dhabi – UAE

### CHINA

**Shanghai**

Building 10, No.401 ChuangQiao Rd.,  
YouDu Space, Pudong District  
Shanghai 201 206

**Guangzhou**

Mirabeau, a Cognizant Digital Business  
85 Huacheng Avenue  
Zhujiang New Town, Guangzhou China

**Dalian City**

4th Floor, No 1 Huixian Yuan,  
Dalian Hi-tech Industrial Zone, Dalian City,  
Liaoning Province (Ascendas Software Park)

### AUSTRALIA

**Melbourne**

15 William Street  
Level 6, Suite 2,  
Melbourne, VIC, 3000

**Sydney**

Level 4  
28 Margaret Street  
Sydney, NSW, 2000

### JAPAN

**Tokyo**

PMO Hanzomon 2-1,  
Kojimachi, Chiyoda-ku,  
Tokyo 102-0083,Japan

### PHILIPPINES

2/F Science Hub Tower 4,  
Campus Ave., Mckinley Hill,  
Fort Bonifacio Taguig City

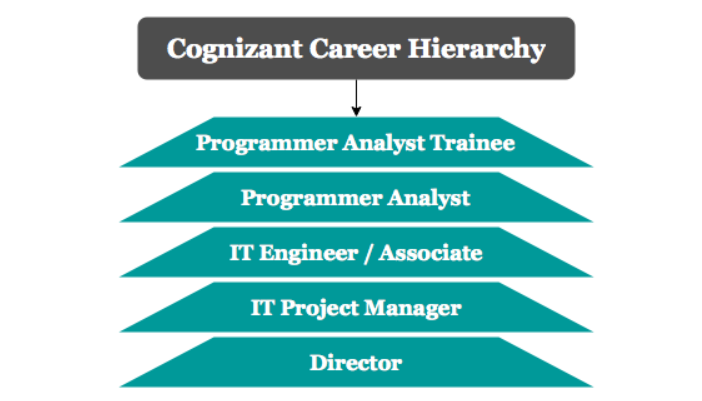
**Divisions of company**

1. **Banking & Financial Services**
2. **Insurance**
3. **Healthcare**
4. **Manufacturing and Retail**
5. **Automotive**
6. **[Communications, Media & Technology](https://www.cognizant.com/content/cognizant/us/en/index/industries/cmt-solutions.html" \o "Communications, Media & Technology)**
7. **Education**
8. **Information Service**
9. **Oil & Gas**
10. **Life Science**
11. **Transportation & Logistics**
12. **Travel & Hospitality**
13. **Utilities**

**My Group and Division**

Big data & Pyspark Developer(IT)

**Administrative tree**



**Main functions of the company**

Cognizant is a global IT services provider, offering consulting and outsourcing services to some of the world's largest enterprises spanning the financial services, media and communications, healthcare, natural resources, and consumer products industries. Cognizant employs nearly 300,000 people globally, roughly 70% of whom are in India, although the company's headquarters are in Teaneck, New Jersey.

Cognizant is organized into several [verticals](https://en.wikipedia.org/wiki/Vertical_market" \o "Vertical market) and [horizontal](https://en.wikipedia.org/wiki/Horizontal_market" \o "Horizontal market) units. The vertical units focus on specific industries such as Banking & Financial Services, Insurance, Healthcare, Manufacturing and Retail. The horizontals focus on specific technologies or process areas such as Analytics, mobile computing, BPO and Testing. Both horizontal and vertical units have [business consultants](https://en.wikipedia.org/wiki/Business_consulting" \o "Business consulting), who together form the organization-wide Cognizant Consulting team. Cognizant is among the largest recruiters of MBAs in the industry; they are involved in [business development](https://en.wikipedia.org/wiki/Business_development" \o "Business development) and [business analysis](https://en.wikipedia.org/wiki/Business_analysis" \o "Business analysis) for IT services projects.

**Cognizant Market Share**

The company's revenue also dipped nearly 5% to Rs 29,682 crore in the fiscal ended March 31, 2021, from Rs 31,196 crore in FY2020, according to a Registrar of Companies filing that was shared by market intelligence firm Tofler.  
  
“The statutory filings in India are not representative of the company’s overall performance," the Nasdaq-listed company said. "Cognizant has reported a record setting second quarter; highest ever quarterly revenue at $4.6 billion; the largest-percentage quarterl ..

CTSI saw its total expenses increasing to Rs 24,981.7 crore in FY21. The employee benefit expense has risen to Rs 19,827 crore in FY21 over the previous fiscal, as per the document.  
  
“During the financial year, the company continues to witness growth owing to an enhanced relationship with company's most trusted customers and initiating new long term partnerships with leading businesses in India spanning across private banks, large pharma and consumer good companies," the filing said.

BENGALURU: Nasdaq-listed Cognizant Technology Solutions Corp’s March quarter net profit rose 11.5% year-on-year to $563 million, on the back of an 11% increase in revenue in constant currency to $4.8 billion - its highest-ever quarterly revenue and fourth consecutive quarter of revenue growth. The company follows a January to December accounting year.

In comparison, for the March quarter, Tata Consultancy Services Ltd’s (TCS) dollar revenue grew 14.3% in constant currency from a year earlier to $6.7 billion, while that of Infosys Ltd grew 20.6% to $4.3 billion.

Teaneck, New Jersey-based Cognizant revised its full year 2022 revenue growth guidance to 9-11% in constant currency at $19.8-20.2 billion which includes an improved organic revenue growth outlook and a lower inorganic contribution, reflecting a disciplined acquisition strategy, the company said. Last quarter, it had projected 8.5-11.5% revenue growth for 2022 to $20.0-20.5 billion, its highest-ever annual revenue outlook.

The company’s digital revenue grew 20% annually and contributed 50% to total revenues for the March quarter, representing continued demand environment as digital technologies become increasingly mainstream. It continues to focus on its “digital battleground" areas of digital engineering, artificial intelligence (AI) and analytics, cloud, and internet of things (IoT).

“While the economic backdrop is uncertain, we remain optimistic about the demand outlook for our solutions. Our strategic repositioning enables us to engage more deeply with clients, helping them succeed, and supports our growth trajectory," said Brian Humphries, chief executive officer, Cognizant.

Revenue from financial services grew 6% in constant currency from a year-ago reflecting demand for digital services, partially offset by clients' focus on cost optimisation. In February, the sale of its Samlink subsidiary was completed, which negatively impacted segment revenue growth by approximately 130 basis points.

**Chapter 2**

**Project Description:**

Real Time Parking Management System On-Demand solution provides a complete set of tools to optimize and make ease of parking vehicle using a ordinary CCTV camera connected through the system. All the information regarding the location of parking can feed by any picture of the parking lodge and than activate the application. In real-time situations, we use the SSD algorithm to get the elements. In addition, the SSD showed results with high reliability. The main purpose of the SSD algorithm is use to detect various objects like in a video sequence in real time and load them in real time live data. This model has demonstrated the positive effectiveness of detection and tracking in professional facilities and can be in some cases we used to detect, track and respond according to the targeted video at reception. This real-time ecosystem analysis can deliver positive results by improving the security, order and usability of any business. In case of a terrorist attack, we constantly increase the work of finding bullets and bullets to set the alarm. This model can also be used with CCTV systems, drones and other surveillance devices to watched in many areas where parking management is difficult, such as schools, government agencies and hospitals. This application simplifies your work by performing fully automatic task management. All functions, such as connecting to the vehicle screen and exchanging real-time data with the driver, make parking management less efficient. The project can help us in making parking management system functionality better and total machine dependent.

**Opportunity :**

Working within an engagement team, a Technical Consultant will be responsible for understanding and managing integration requirements for cloud systems, technical architecture, prototyping, process design, testing, cut over, training, defining support procedures, leading work streams and supporting implementations

●    Estimation of development, deployment and testing of new requirements or modifications in existing solutions.

●    Implement and deploy custom solutions, integration solutions and ensure best practices & standards have been followed

●    Conduct technical design sessions with clients to propose architectural choices and create detailed technical documents, like technical specs, system architecture diagrams, data models etc.

●     Troubleshoot issues that arise during unit, system integration and user acceptance testing

●      Provide on-going post implementation support to optimize user adoption

The Work done during the internship

●      Learning new technology like Hadoop and its environment, Pyspark, Linux, Python and Scala on given time interval.

●      Business Training to help in giving proper industrial knowledge of soft skills

●       Good written and oral communication skills.

●  Experience working with other teams, Experience in communication and documentation.

Reason for choosing location: Known to be one of the firsts to get the top four IT companies to set up centres on its turf, Noida is sensing an opportunity to be the hub for accommodating the next phase of ramp-up by information technology (IT) and ITes (IT enabled services). The state government is keen to rope in IT and BPO (Business Process Outsourcing) firms who want to expand footprint, but are mindful of the costs of organic growth. As a destination, Noida offers both skilled and an industry ready workforce with a cost advantage not matched by many.

Cognizant prefer Data Science, Data Analytics and CDC in location Noida. Customer centricity is at the core of Cognizant Noida Office in Noida Sector 135, Noida, Delhi and it is this belief that has led the business to build long-term relationships. Ensuring a positive customer experience, making available goods and/or services that are of top-notch quality is given prime importance.

India’s leading B2B market place, Jd Mart ensures engaging in business activities is a seamless process for small and medium enterprises as well as large businesses. In a wake to enable these businesses to reach their audience, this portal lets them showcase their offerings in terms of the products and/or services through a digital catalogue. This business has a wide range of product offerings and the product/catalogue list includes Corporate Companies etc.

**Chapter 3**

**Technology Details**

**Advance Python :**  Python is a computer programming language often used to build websites and software, automate tasks, and conduct data analysis. Python is a general-purpose language, meaning it can be used to create a variety of different programs and isn’t specialized for any specific problems. This versatility, along with its beginner-friendliness, has made it one of the most-used programming languages today. It uses a simplified syntax with an emphasis on natural language, for a much easier learning curve for beginners. And, because Python is free to use and is supported by an extremely large ecosystem of libraries and packages, it's often the first-choice language for new developers.

Usage of Python-:

### 1. AI and machine learning

Because Python is such a stable, flexible, and simple programming language, it’s perfect for various machine learning (ML) and artificial intelligence (AI) projects. In fact, Python is among the favourite languages among data scientists, and there are many Python machine learning and AI libraries and packages available.

2. Data analytics

Much like AI and machine learning, data analytics is another rapidly developing field that utilises Python programming. At a time when we’re creating more data than ever before, there is a need for those who can collect, manipulate and organise the information.

Python for data science and analytics makes sense. The language is easy-to-learn, flexible, and well-supported, meaning it’s relatively quick and easy to use for analysing data. When working with large amounts of information, it’s useful for manipulating data and carrying out repetitive tasks.

### 3. Data visualisation

Data visualisation is another popular and developing area of interest. Again, it plays into many of the strengths of Python. As well as its flexibility and the fact it’s open-source, Python provides a variety of graphing libraries with all kinds of features.

### 4. Programming applications

You can program all kinds of applications using Python. The general-purpose language can be used to read and create file directories, create GUIs and APIs, and more. Whether it’s blockchain applications, audio and video apps, or machine learning applications, you can build them all with Python.

### 5. Web development

Python is a great choice for web development. This is largely due to the fact that there are many Python web development frameworks to choose from, such as [Django](https://www.futurelearn.com/courses/building-web-applications-in-django), Pyramid, and Flask. These frameworks have been used to create sites and services such as Spotify, Reddit and Mozilla.

### 6. Game development

Although far from an industry-standard in game development, Python does have its uses in the industry. It’s possible to create simple games using the programming language, which means it can be a useful tool for quickly developing a prototype. Similarly, certain functions (such as dialogue tree creation) are possible in Python.

### 7. Language development

The simple and elegant design of Python and its syntax means that it has inspired the creation of new programming languages. Languages such as Cobra, CoffeeScript, and Go all use a similar syntax to Python.

### 8. Finance

Python is increasingly being utilised in the world of finance, often in areas such as quantitative and qualitative analysis. It can be a valuable tool in determining asset price trends and predictions, as well as in automating workflows across different data sources.

### 9. SEO

Another slightly surprising entry on our list of Python uses is in the field of [search engine optimisation (SEO)](https://www.futurelearn.com/courses/the-ultimate-seo-wordpress-training-course). It’s an area that often benefits from automation, which is certainly possible through Python. Whether it’s implementing changes across multiple pages or categorising keywords, Python can help.

### 10. Design

When asking ‘what is Python used for?’ you probably weren’t expecting design to feature on the list. However, Python can be used to develop graphic design applications. Surprisingly, the language is used across a range of 2D imaging software, such as Paint Shop Pro and Gimp.

**Big Data -:**

Big data is a combination of structured, semistructured and unstructured data collected by organizations that can be mined for information and used in [machine learning](https://www.techtarget.com/searchenterpriseai/definition/machine-learning-ML) projects, [predictive modeling](https://www.techtarget.com/searchenterpriseai/definition/predictive-modeling) and other advanced analytics applications.

Systems that process and store big data have become a common component of [data management](https://www.techtarget.com/searchdatamanagement/definition/data-management) architectures in organizations, combined with tools that support [big data analytics](https://www.techtarget.com/searchbusinessanalytics/definition/big-data-analytics) uses. Big data is often characterized by the three V's:

**(i) Volume –** The name Big Data itself is related to a size which is enormous. Size of data plays a very crucial role in determining value out of data. Also, whether a particular data can actually be considered as a Big Data or not, is dependent upon the volume of data. Hence, ****‘Volume’**** is one characteristic which needs to be considered while dealing with Big Data solutions.

**(ii) Variety –** The next aspect of Big Data is its ****variety****.

Variety refers to heterogeneous sources and the nature of data, both structured and unstructured. During earlier days, spreadsheets and databases were the only sources of data considered by most of the applications. Nowadays, data in the form of emails, photos, videos, monitoring devices, PDFs, audio, etc. are also being considered in the analysis applications. This variety of unstructured data poses certain issues for storage, mining and analyzing data.

**(iii) Velocity –** The term ****‘velocity’**** refers to the speed of generation of data. How fast the data is generated and processed to meet the demands, determines real potential in the data.

Big Data Velocity deals with the speed at which data flows in from sources like business processes, application logs, networks, and social media sites, sensors,[Mobile](https://www.guru99.com/mobile-testing.html)devices, etc. The flow of data is massive and continuous.

**(iv) Variability –** This refers to the inconsistency which can be shown by the data at times, thus hampering the process of being able to handle and manage the data effectively.

## **Types Of Big Data**

Following are the types of Big Data:

1. ****Structured****
2. ****Unstructured****
3. ****Semi-structured****

### ****Structured****

Any data that can be stored, accessed and processed in the form of fixed format is termed as a ‘structured’ data. Over the period of time, talent in computer science has achieved greater success in developing techniques for working with such kind of data (where the format is well known in advance) and also deriving value out of it. However, nowadays, we are foreseeing issues when a size of such data grows to a huge extent, typical sizes are being in the rage of multiple zettabytes.

### ****Unstructured****

Any data with unknown form or the structure is classified as unstructured data. In addition to the size being huge, un-structured data poses multiple challenges in terms of its processing for deriving value out of it. A typical example of unstructured data is a heterogeneous data source containing a combination of simple text files, images, videos etc. Now day organizations have wealth of data available with them but unfortunately, they don’t know how to derive value out of it since this data is in its raw form or unstructured format.

****Examples -:****  ‘Google Search’

### ****Semi-structured****

Semi-structured data can contain both the forms of data. We can see semi-structured data as a structured in form but it is actually not defined with e.g. a table definition in relational [DBMS](https://www.guru99.com/what-is-dbms.html). Example of semi-structured data is a data represented in an XML file.

**Examples** -: Personal data stored in an XML file

## **Advantages Of Big Data Processing**

Ability to process Big Data in DBMS brings in multiple benefits, such as-

* Businesses can utilize outside intelligence while taking decisions

Access to social data from search engines and sites like facebook, twitter are enabling organizations to fine tune their business strategies.

* Improved customer service

Traditional customer feedback systems are getting replaced by new systems designed with Big Data technologies. In these new systems, Big Data and natural language processing technologies are being used to read and evaluate consumer responses.

* Early identification of risk to the product/services, if any
* Better operational efficiency

Big Data technologies can be used for creating a staging area or landing zone for new data before identifying what data should be moved to the [data warehouse](https://www.guru99.com/data-warehousing.html). In addition, such integration of Big Data technologies and data warehouse helps an organization to offload infrequently accessed data.

**Hadoop -:**

Apache Hadoop is an open source framework that is used to efficiently store and process large datasets ranging in size from gigabytes to petabytes of data. Instead of using one large computer to store and process the data, Hadoop allows clustering multiple computers to analyze massive datasets in parallel more quickly.

With the advent of the World Wide Web in the late 1990s and early 2000s, search engines and indexes were created to help locate relevant information within textual content. Initially, search results were returned by humans. Of course, when the number of pages increased to tens of millions, automation became necessary. Web crawlers were then created, mainly as university research projects. Search engines like Yahoo and AltaVista have also started to appear.

Among these search engines, the open source project Nutch was created by Doug Cutting and Mike Cafarella. Their goal was to deliver web search results faster by distributing data and calculations across different computers to accomplish multiple tasks simultaneously. At the same time, the Google search engine was in development. This project was based on the same concept of storing and processing data in a distributed and automated way to deliver search results faster.

## How is Hadoop used by businesses?

Today, beyond its initial functionality to search millions of web pages for relevant results, Hadoop is used by many companies as a big data platform. Here are the main uses of Hadoop in business today:

**Low-cost storage and data archiving**

The modest cost of standard machines makes this data processing platform very useful for storing and combining data. Transactional data, or data coming from social networks, machines, scientific data, click streams… low-cost storage makes it possible to keep information that is not particularly useful at the moment in case it is needed. would become so later.

**A sandbox for discovery and analysis**

Designed to process large volumes of data of different shapes, Hadoop is capable of handling analytical algorithms. Analytical tools can help the business operate more efficiently, discover new opportunities, and gain competitive advantages. Hadoop's sandbox approach offers opportunities for medium innovation with minimal investment.

**Data Lakes**

Data Lakes support storing data in the original format. The goal is to provide a raw, unrefined view of data for data scientists and analysts for discovery and analysis. This helps them to ask new or complex questions without constraints. Data Lakes are not, however, a replacement for Data Warehouses. How to secure and govern Data Lakes remains a big debate in the IT industry today. It may be necessary to develop logical data structures based on data federation techniques.

**Complement the Data Warehouses**

Currently, Hadoop sits alongside Data Warehouse environments. Likewise, some datasets are offloaded directly from Data Warehouses to Hadoop, and some new types of data go directly to Hadoop. The end goal of every business is to have the right platform to store and process data of different schemas and formats to support different use cases that can be integrated at different levels.

**Hadoop and the Internet of Things**

Connected objects need to know what to communicate and when to act. The essence of the Internet of Things is the continuous streaming of data. Hadoop is often used as a data store for billions of transactions. Massive storage and processing capabilities also allow the big data platform to be used as a sandbox for discovering and establishing patterns for prescriptive instruction. These guidelines can then be continuously improved, as Hadoop is constantly updated with new data.

**Recommendations engine**

One of the most popular uses of Hadoop is for building recommendation systems on the web. Many companies use the framework's analytical tools to provide these types of services. Facebook uses it to suggest people you might know, LinkedIn offers you jobs that you might be interested in, and Netflix, eBay, and Hulu recommend content. These recommendation systems analyze large amounts of data in real time to quickly predict consumer preferences before they have time to leave the web page.

Hadoop consists of four main modules:

**Hadoop Distributed File System (HDFS)** – A distributed file system that runs on standard or low-end hardware. HDFS provides better data throughput than traditional file systems, in addition to high fault tolerance and native support of large datasets.

**example of HDFS**

Consider a file that includes the phone numbers for everyone in the United States; the numbers for people with a last name starting with A might be stored on server 1, B on server 2, and so on.

With Hadoop, pieces of this phonebook would be stored across the cluster, and to reconstruct the entire phonebook, your program would need the blocks from every server in the cluster.

To ensure availability if and when a server fails, HDFS replicates these smaller pieces onto two additional servers by default. (The redundancy can be increased or decreased on a per-file basis or for a whole environment; for example, a development Hadoop cluster typically doesn’t need any data redundancy.) This redundancy offers multiple benefits, the most obvious being higher availability.

The redundancy also allows the Hadoop cluster to break up work into smaller chunks and run those jobs on all the servers in the cluster for better scalability. Finally, you gain the benefit of data locality, which is critical when working with large data sets.

**Yet Another Resource Negotiator (YARN)** – Manages and monitors cluster nodes and resource usage. It schedules jobs and tasks.

## Why YARN?

MapReduce performs functions of Resource Management and Processing. Hadoop v1.0 is also known as MapReduce Version 1 (MRV1). There was only a single master for Job Tracker.

In the previous version of Hadoop that is Hadoop version 1.0, which is also known as MapReduce version 1 (MRV1) use to perform both the task of process and resource management by itself. It has a job tracker module that is responsible for everything. Hence it is the single master that allocates resources for applications, performs scheduling for demand and also monitors the jobs of processing in the system. Hadoop version 1.0 reduces tasks & assigns maps on several sub-processes which is called Task Trackers. Task Tracker also reports the progress of processes in a periodical manner. But the main issue is not that, the problem is this design of a single master for all, resulting in bottlenecking issue. Also, the computational resource utilization was inefficient. Thus scalability became an issue with this version of Hadoop. But on the bright side, this issue is resolved by YARN, a vital core component in its successor Hadoop version 2.0 which was introduced in the year 2012 by Yahoo and Hortonworks. The basic idea behind this relief is separating MapReduce from Resource Management and Job scheduling instead of a single master. Thus, YARN is now responsible for Job scheduling and Resource Management.

 In Hadoop 2.0, The concept of Application Master and Resource Manager was introduced by YARN. Across the cluster of Hadoop, the utilization of resources is monitored by the Resource Manager.

****There are some features of YARN because of which it got very famous, which are:****

1. **Multi-tenancy:** YARN has allowed access to multiple data processing engines such as batch processing engine, stream processing engine, interactive processing engine, graph processing engine and much more. This has given the benefit of multi-tenancy to the company.
2. **Cluster Utilization:** Clusters are utilized in an optimized way because clusters are used dynamically in Hadoop with the help of YARN.
3. **Compatibility**: YARN is also compatible with the first version of Hadoop, i.e. Hadoop 1.0, because it uses the existing map-reduce apps. So YARN can also be used with Hadoop 1.0.
4. **Scalability**: Thousands of clusters and nodes are allowed by the scheduler in Resource Manager of YARN to be managed and extended by Hadoop.

**MapReduce** – A framework that helps programs do the parallel computation on data. The map task takes input data and converts it into a dataset that can be computed in key value pairs. The output of the map task is consumed by reduce tasks to aggregate output and provide the desired result.

****MapReduce programming offers several benefits to help you gain valuable insights from your big data:****

* ****Scalability****. Businesses can process petabytes of data stored in the Hadoop Distributed File System (HDFS).
* ****Flexibility****. Hadoop enables easier access to multiple sources of data and multiple types of data.
* ****Speed****. With parallel processing and minimal data movement, Hadoop offers fast processing of massive amounts of data.
* ****Simple****. Developers can write code in a choice of languages, including Java, C++ and Python.

**Example-:**

Using the MapReduce framework, you can break this down into five map tasks, where each mapper works on one of the five files. The mapper task goes through the data and returns the maximum temperature for each city.

For example, the results produced from one mapper task for the data above would look like this: (Toronto, 20) (Whitby, 25) (New York, 22) (Rome, 33)

Assume the other four mapper tasks (working on the other four files not shown here) produced the following intermediate results:

(Toronto, 18) (Whitby, 27) (New York, 32) (Rome, 37) (Toronto, 32) (Whitby, 20) (New York, 33) (Rome, 38) (Toronto, 22) (Whitby, 19) (New York, 20) (Rome, 31) (Toronto, 31) (Whitby, 22) (New York, 19) (Rome, 30)

All five of these output streams would be fed into the reduce tasks, which combine the input results and output a single value for each city, producing a final result set as follows: (Toronto, 32) (Whitby, 27) (New York, 33) (Rome, 38).

**Hadoop Common** – Provides common Java libraries that can be used across all modules. The Hadoop Common package is considered as the base/core of the framework as it provides essential services and basic processes such as abstraction of the underlying operating system and its file system. Hadoop Common also contains the necessary Java Archive (JAR) files and scripts required to start Hadoop. The Hadoop Common package also provides source code and documentation, as well as a contribution section that includes different projects from the Hadoop Community.

Apart from these four main modules, various components based on the framework have achieved a high level of reputation among Apache projects:

**Ambari** is a web interface to manage, configure and test Hadoop services and components.

Some of the key features of this technology can be highlighted as:

1. Instantaneous insight into the health of the Hadoop cluster using preconfigured operational metrics
2. User-friendly configuration providing an easy step-by-step guide for installation
3. Installation of Apache Ambari is possible through Hortonworks Data Platform (HDP)
4. Monitoring dependencies and performances by visualizing and analyzing jobs and tasks
5. Authentication, authorization, and auditing by installing Kerberos-based Hadoop clusters
6. Flexible and adaptive technology fitting perfectly in the enterprise environment

## ****Installing Apache Ambari****

To build up the cluster, the Install Wizard needs to know some general information regarding the cluster to which you should supply the fully qualified domain name (FQDN) of your each host.

Additionally, the wizard needs access to the private key file the user created in Set Up Passwordless SSH. This is used to locate all the hosts in the system and to access and interact with them securely.

1. The list of hostnames, one per line, can be entered using the **Target Hosts**text box

2. Select **Provide Your SSH Private Key**if you want Ambari to automatically install the Ambari Agent on all your hosts using SSH. In the **Host Registration Information**, you can use the **Choose File** button to find the private key file matching the public key installed earlier on all your hosts. Alternatively, you can cut and paste the key into the text box manually

3. Select **Perform Manual Registration** if you do not wish Ambari to automatically install the Ambari Agent

**Flume** is software that collects, aggregates and moves large amounts of data streams under HDFS.

1. In the above diagram, the events generated by external source (WebServer) are consumed by Flume Data Source. The external source sends events to Flume source in a format that is recognized by the target source.
2. Flume Source receives an event and stores it into one or more channels. The channel acts as a store which keeps the event until it is consumed by the flume sink. This channel may use a local file system in order to store these events.
3. Flume sink removes the event from a channel and stores it into an external repository like e.g., HDFS. There could be multiple flume agents, in which case flume sink forwards the event to the flume source of next flume agent in the flow.

## **Some Important features of FLUME**

1. Flume has a flexible design based upon streaming data flows. It is fault tolerant and robust with multiple failovers and recovery mechanisms. Flume Big data has different levels of reliability to offer which includes **‘best-effort delivery’** and an **‘end-to-end delivery’**. **Best-effort delivery** does not tolerate any Flume node failure whereas **‘end-to-end delivery’** mode guarantees delivery even in the event of multiple node failures.
2. Flume carries data between sources and sinks. This gathering of data can either be scheduled or event-driven. Flume has its own query processing engine which makes it easy to transform each new batch of data before it is moved to the intended sink.
3. Possible **Flume sinks** include **HDFS** and **HBase**. Flume Hadoop can also be used to transport event data including but not limited to network traffic data, data generated by social media websites and email messages.

**HBase** is a distributed, non-relational database that runs on top of Hadoop. Its arrays can be used as input and output for MapReduce tasks.

## Limitations of Hadoop

Hadoop can perform only batch processing, and data will be accessed only in a sequential manner. That means one has to search the entire dataset even for the simplest of jobs.

A huge dataset when processed results in another huge data set, which should also be processed sequentially. At this point, a new solution is needed to access any point of data in a single unit of time (random access).

## example HBase

An HBase column represents an attribute of an object; if the table is storing diagnostic logs from servers in your environment, each row might be a log record, and a typical column  could be the timestamp of when the log record was written, or the server name where the record originated.

HBase allows for many attributes to be grouped together into column families, such that the elements of a column family are all stored together. This is different from a row-oriented relational database, where all the columns of a given row are stored together. With HBase you must predefine the table schema and specify the column families. However, new columns can be added to families at any time, making the schema flexible and able to adapt to changing application requirements.

Just as HDFS has a NameNode and slave nodes, and MapReduce has JobTracker and TaskTracker slaves, HBase is built on similar concepts. In HBase a master node manages the cluster and region servers store portions of the tables and perform the work on the data. In the same way HDFS has some enterprise concerns due to the availability of the NameNode HBase is also sensitive to the loss of its master node.

**Hive** is a data warehouse and SQL query language presenting data in tabular form. Programming Hive is similar to programming a database. Hive is a data warehouse system which is used to analyze structured data. It is built on the top of Hadoop. It was developed by Facebook. Hive provides the functionality of reading, writing, and managing large datasets residing in distributed storage. It runs SQL like queries called HQL (Hive query language) which gets internally converted to MapReduce jobs. Using Hive, we can skip the requirement of the traditional approach of writing complex MapReduce programs. Hive supports Data Definition Language (DDL), Data Manipulation Language (DML), and User Defined Functions (UDF).

**Features of Hive**

* Hive is fast and scalable.
* It provides SQL-like queries (i.e., HQL) that are implicitly transformed to MapReduce or Spark jobs.
* It is capable of analyzing large datasets stored in HDFS.
* It allows different storage types such as plain text, RCFile, and HBase.
* It uses indexing to accelerate queries.
* It can operate on compressed data stored in the Hadoop ecosystem.
* It supports user-defined functions (UDFs) where user can provide its functionality.

## Limitations of Hive

* Hive is not capable of handling real-time data.
* It is not designed for online transaction processing.
* Hive queries contain high latency.

**Pig** is a platform for manipulating data stored in HDFS, and includes a compiler for MapReduce programs as well as a high-level language called Pig Latin. It provides a way to perform data extractions, transformations and loads, as well as basic analysis without having to write MapReduce programs.

**Evolution of Pig:**Earlier in 2006, Apache Pig was developed by Yahoo’s researchers. At that time, the main idea to develop Pig was to execute the MapReduce jobs on extremely large datasets. In the year 2007, it moved to Apache Software Foundation(ASF) which makes it an open source project. The first version(*0.1*) of Pig came in the year 2008. The latest version of Apache Pig is *0.18* which came in the year 2017.

**Features of Apache Pig:**

1. For performing several operations Apache Pig provides rich sets of operators like the filters, join, sort, etc.
2. Easy to learn, read and write. Especially for SQL-programmer, Apache Pig is a boon.
3. Apache Pig is extensible so that you can make your own user-defined functions and process.
4. Join operation is easy in Apache Pig.
5. Fewer lines of code.
6. Apache Pig allows splits in the pipeline.
7. The data structure is multivalued, nested, and richer.
8. Pig can handle the analysis of both structured and unstructured data.

**Applications of Apache Pig:**

1. For exploring large datasets Pig Scripting is used.
2. Provides the supports across large data-sets for Ad-hoc queries.
3. In the prototyping of large data-sets processing algorithms.
4. Required to process the time sensitive data loads.
5. For collecting large amounts of datasets in form of search logs and web crawls.
6. Used where the analytical insights are needed using the sampling.

**Types of Data Models in Apache Pig:** It consist of the 4 types of data models as follows:

1. **Atom**: It is a atomic data value which is used to store as a string. The main use of this model is that it can be used as a number and as well as a string.
2. **Tuple**: It is an ordered set of the fields.
3. **Bag**: It is a collection of the tuples.
4. **Map**: It is a set of key/value pairs.

**Hadoop Spark** is an open-source cluster computing framework with in-memory analytics tools

Apache Spark has following features.

1. **Speed** − Spark helps to run an application in Hadoop cluster, up to 100 times faster in memory, and 10 times faster when running on disk. This is possible by reducing number of read/write operations to disk. It stores the intermediate processing data in memory.
2. **Supports multiple languages** − Spark provides built-in APIs in Java, Scala, or Python. Therefore, you can write applications in different languages. Spark comes up with 80 high-level operators for interactive querying.
3. **Advanced Analytics** − Spark not only supports ‘Map’ and ‘reduce’. It also supports SQL queries, Streaming data, Machine learning (ML), and Graph algorithms.

There are three ways of Spark deployment as explained below.

1. **Standalone** − Spark Standalone deployment means Spark occupies the place on top of HDFS(Hadoop Distributed File System) and space is allocated for HDFS, explicitly. Here, Spark and MapReduce will run side by side to cover all spark jobs on cluster.
2. **Hadoop Yarn** − Hadoop Yarn deployment means, simply, spark runs on Yarn without any pre-installation or root access required. It helps to integrate Spark into Hadoop ecosystem or Hadoop stack. It allows other components to run on top of stack.
3. **Spark in MapReduce (SIMR)** − Spark in MapReduce is used to launch spark job in addition to standalone deployment. With SIMR, user can start Spark and uses its shell without any administrative access.

**Component of Spark**

### Apache Spark Core

Spark Core is the underlying general execution engine for spark platform that all other functionality is built upon. It provides In-Memory computing and referencing datasets in external storage systems.

### Spark SQL

Spark SQL is a component on top of Spark Core that introduces a new data abstraction called SchemaRDD, which provides support for structured and semi-structured data.

### Spark Streaming

Spark Streaming leverages Spark Core's fast scheduling capability to perform streaming analytics. It ingests data in mini-batches and performs RDD (Resilient Distributed Datasets) transformations on those mini-batches of data.

### MLlib (Machine Learning Library)

MLlib is a distributed machine learning framework above Spark because of the distributed memory-based Spark architecture. It is, according to benchmarks, done by the MLlib developers against the Alternating Least Squares (ALS) implementations. Spark MLlib is nine times as fast as the Hadoop disk-based version of **Apache Mahout** (before Mahout gained a Spark interface).

### GraphX

GraphX is a distributed graph-processing framework on top of Spark. It provides an API for expressing graph computation that can model the user-defined graphs by using Pregel abstraction API. It also provides an optimized runtime for this abstraction.

**Apache Nifi**

Apache NiFi is an integrated data logistics platform for automating the movement of data between disparate systems. It provides real-time control that makes it easy to manage the movement of data between any source and any destination. It is data source agnostic, supporting disparate and distributed sources of differing formats, schemas, protocols, speeds and sizes such as machines, geo location devices, click streams, files, social feeds, log files and videos and more. It is configurable plumbing for moving data around, similar to how Fedex, UPS or other courier delivery services move parcels around. And just like those services, Apache NiFi allows you to trace your data in real time, just like you could trace a delivery.

Apache NiFi is based on technology previously called “Niagara Files” that was in development and used at scale within the NSA for the last eight years and was made available to the Apache Software Foundation through the NSA Technology Transfer Program. As such, it was designed from the beginning to be field ready—flexible, extensible and suitable for a wide range of devices from a small lightweight network edge device such as a Raspberry Pi to enterprise data clusters and the cloud. Apache NiFi is also able to dynamically adjust to fluctuating network connectivity that could impact communications and thus the delivery of data.

**Sqoop** is a connection and transfer mechanism for moving data between Hadoop and relational databases.

**Some of the important Features of the Sqoop :**

1. Sqoop also helps us to connect the result from the SQL Queries into Hadoop distributed file system.
2. Sqoop helps us to load the processed data directly into the hive or Hbase.
3. It performs the security operation of data with the help of Kerberos.
4. With the help of Sqoop, we can perform compression of processed data.
5. Sqoop is highly powerful and efficient in nature.

**There are two major operations performed in Sqoop :**

1. Import
2. Export

**Sqoop Import :**

Sqoop import command helps in implementation of the operation. With the help of the import command, we can import a table from the Relational database management system to the Hadoop database server. Records in Hadoop structure are stored in text files and each record is imported as a separate record in Hadoop database server. We can also create load and partition in Hive while importing data..Sqoop also supports incremental import of data which means in case we have imported a database and we want to add some more rows, so with the help of these functions we can only add the new rows to existing database, not the complete database.

**Sqoop Export :**

Sqoop export command helps in the implementation of operation. With the help of the export command which works as a reverse process of operation. Herewith the help of the export command we can transfer the data from the Hadoop database file system to the Relational database management system. The data which will be exported is processed into records before operation is completed. The export of data is done with two steps, first is to examine the database for metadata and second step involves migration of data.

**Zookeeper** is an application to coordinate distributed treatments.

### What ZooKeeper Does

ZooKeeper provides a very simple interface and services. ZooKeeper brings these key benefits:

1. Fast. ZooKeeper is especially fast with workloads where reads to the data are more common than writes. The ideal read/write ratio is about 10:1.
2. Reliable. ZooKeeper is replicated over a set of hosts (called an ensemble) and the servers are aware of each other. As long as a critical mass of servers is available, the ZooKeeper service will also be available. There is no single point of failure.
3. Simple. ZooKeeper maintain a standard hierarchical name space, similar to files and directories.
4. Ordered. The service maintains a record of all transactions, which can be used for higher-level abstractions, like synchronization primitives.

## **How ZooKeeper Works**

ZooKeeper allows distributed processes to coordinate with each other through a shared hierarchical name space of data registers, known as znodes. Every znode is identified by a path, with path elements separated by a slash (“/”). Aside from the root, every znode has a parent, and a znode cannot be deleted if it has children.

This is much like a normal file system, but ZooKeeper provides superior reliability through redundant services. A service is replicated over a set of machines and each maintains an in-memory image of the the data tree and transaction logs. Clients connect to a single ZooKeeper server and maintains a TCP connection through which they send requests and receive responses.

This architecture allows ZooKeeper to provide high throughput and availability with low latency, but the size of the database that ZooKeeper can manage is limited by memory.

**Apache Kafka -:**

#### [What is event streaming?](https://kafka.apache.org/intro" \l "intro_streaming)

Event streaming is the digital equivalent of the human body's central nervous system. It is the technological foundation for the 'always-on' world where businesses are increasingly software-defined and automated, and where the user of software is more software.

Technically speaking, event streaming is the practice of capturing data in real-time from event sources like databases, sensors, mobile devices, cloud services, and software applications in the form of streams of events; storing these event streams durably for later retrieval; manipulating, processing, and reacting to the event streams in real-time as well as retrospectively; and routing the event streams to different destination technologies as needed. Event streaming thus ensures a continuous flow and interpretation of data so that the right information is at the right place, at the right time.

#### [What can I use event streaming for?](https://kafka.apache.org/intro" \l "intro_usage)

Event streaming is applied to a [wide variety of use cases](https://kafka.apache.org/powered-by) across a plethora of industries and organizations. Its many examples include:

* To process payments and financial transactions in real-time, such as in stock exchanges, banks, and insurances.
* To track and monitor cars, trucks, fleets, and shipments in real-time, such as in logistics and the automotive industry.
* To continuously capture and analyze sensor data from IoT devices or other equipment, such as in factories and wind parks.
* To collect and immediately react to customer interactions and orders, such as in retail, the hotel and travel industry, and mobile applications.
* To monitor patients in hospital care and predict changes in condition to ensure timely treatment in emergencies.
* To connect, store, and make available data produced by different divisions of a company.
* To serve as the foundation for data platforms, event-driven architectures, and microservices.

#### [Apache Kafka® is an event streaming platform. What does that mean?](https://kafka.apache.org/intro" \l "intro_platform)

Kafka combines three key capabilities so you can implement [your use cases](https://kafka.apache.org/powered-by) for event streaming end-to-end with a single battle-tested solution:

1. To **publish** (write) and **subscribe to** (read) streams of events, including continuous import/export of your data from other systems.
2. To **store** streams of events durably and reliably for as long as you want.
3. To **process** streams of events as they occur or retrospectively.

And all this functionality is provided in a distributed, highly scalable, elastic, fault-tolerant, and secure manner. Kafka can be deployed on bare-metal hardware, virtual machines, and containers, and on-premises as well as in the cloud. You can choose between self-managing your Kafka environments and using fully managed services offered by a variety of vendors.

#### [How does Kafka work in a nutshell?](https://kafka.apache.org/intro" \l "intro_nutshell)

Kafka is a distributed system consisting of **servers** and **clients** that communicate via a high-performance [TCP network protocol](https://kafka.apache.org/protocol.html). It can be deployed on bare-metal hardware, virtual machines, and containers in on-premise as well as cloud environments.

**Servers**: Kafka is run as a cluster of one or more servers that can span multiple datacenters or cloud regions. Some of these servers form the storage layer, called the brokers. Other servers run [Kafka Connect](https://kafka.apache.org/documentation/" \l "connect) to continuously import and export data as event streams to integrate Kafka with your existing systems such as relational databases as well as other Kafka clusters. To let you implement mission-critical use cases, a Kafka cluster is highly scalable and fault-tolerant: if any of its servers fails, the other servers will take over their work to ensure continuous operations without any data loss.

**Clients**: They allow you to write distributed applications and microservices that read, write, and process streams of events in parallel, at scale, and in a fault-tolerant manner even in the case of network problems or machine failures. Kafka ships with some such clients included, which are augmented by [dozens of clients](https://cwiki.apache.org/confluence/display/KAFKA/Clients) provided by the Kafka community: clients are available for Java and Scala including the higher-level [Kafka Streams](https://kafka.apache.org/documentation/streams/) library, for Go, Python, C/C++, and many other programming languages as well as REST APIs.

**Pyspark -:**

PySpark is the Python API for Apache Spark, an open source, distributed computing framework  and set of libraries for real-time, large-scale data processing. If you’re already familiar with Python and libraries such as Pandas, then PySpark is a good language to learn to create more scalable analyses and pipelines.

[Apache Spark](https://www.dominodatalab.com/data-science-dictionary/apache-spark/) is basically a computational engine that works with huge sets of data by processing them in parallel and batch systems. Spark is written in Scala, and PySpark was released to support the collaboration of Spark and Python. In addition to providing an API for Spark, PySpark helps you interface with Resilient Distributed Datasets (RDDs) by leveraging the Py4j library.

### **Key Features of PySpark**

* ****Real-time computations****: Because of the in-memory processing in the PySpark framework, it shows low latency.
* ****Polyglot****: The PySpark framework is compatible with various languages such as Scala, Java, Python, and R, which makes it one of the most preferable frameworks for processing huge datasets.
* ****Caching and disk persistence****: This framework provides powerful caching and great disk persistence.
* ****Fast processing****: The PySpark framework is way faster than other traditional frameworks for Big Data processing.
* ****Works well with RDDs****: Python programming language is dynamically typed, which helps when working with RDDs.

### **Spark with Python vs Spark with Scala**

As it is already discussed, Python is not the only programming language that can be used with Apache Spark. Data Scientists already prefer Spark because of the several benefits it has over other Big Data tools, but choosing which language to use with Spark is a dilemma that they face.

Being one of the most popular frameworks when it comes to [Big Data Analytics](https://intellipaat.com/blog/big-data-analytics/" \t "https://intellipaat.com/blog/tutorial/spark-tutorial/pyspark-tutorial/_blank), Python has gained so much popularity that you wouldn’t be shocked if it became the de-facto framework for evaluating and dealing with large datasets and[Machine Learning](https://intellipaat.com/blog/what-is-machine-learning/" \t "https://intellipaat.com/blog/tutorial/spark-tutorial/pyspark-tutorial/_blank) in the coming years.

The most used programming languages with Spark are Python and Scala. Now if you are going to learn PySpark (Spark with Python), then it is important that you know why and when to use Spark with Python, instead of Spark with Scala. In this section, the basic criteria, one should keep in mind while making the choice between Python and Scala to work on Apache Spark, are explained.

**NoSql -:**

NoSQL databases (aka "not only SQL") are non-tabular databases and store data differently than relational tables. NoSQL databases come in a variety of types based on their data model. The main types are document, key-value, wide-column, and graph. They provide flexible schemas and scale easily with large amounts of data and high user loads.

While a variety of differences exist between relational database management systems (RDBMS) and NoSQL databases, one of the key differences is the way the data is modeled in the database. In this section, we'll work through an example of modeling the same data in a relational database and a NoSQL database. Then, we'll highlight some of the other key differences between relational databases and NoSQL databases.

### Types of NoSQL databases

Over time, four major [types of NoSQL databases](https://www.mongodb.com/scale/types-of-nosql-databases" \t "https://www.mongodb.com/_target) emerged: document databases, key-value databases, wide-column stores, and graph databases.

* **Document databases** store data in documents similar to JSON (JavaScript Object Notation) objects. Each document contains pairs of fields and values. The values can typically be a variety of types including things like strings, numbers, booleans, arrays, or objects.
* **Key-value databases** are a simpler type of database where each item contains keys and values.
* **Wide-column stores** store data in tables, rows, and dynamic columns.
* **Graph databases** store data in nodes and edges. Nodes typically store information about people, places, and things, while edges store information about the relationships between the nodes.

### Difference between RDBMS and NoSQL databases

While a variety of differences exist between relational database management systems (RDBMS) and NoSQL databases, one of the key differences is the way the data is modeled in the database. In this section, we'll work through an example of modeling the same data in a relational database and a NoSQL database. Then, we'll highlight some of the other key differences between relational databases and NoSQL databases.

**Linux -:**

Linux is an open-source operating system like other operating systems such as Microsoft [Windows](https://www.javatpoint.com/windows), Apple Mac OS, iOS, Google android, etc. An operating system is a software that enables the communication between computer hardware and software. It conveys input to get processed by the processor and brings output to the hardware to display it. This is the basic function of an operating system. Although it performs many other important tasks, let's not talk about that.

Linux is around us since the mid-90s. It can be used from wristwatches to supercomputers. It is everywhere in our phones, laptops, PCs, cars and even in refrigerators. It is very much famous among developers and normal computer users.

## What's a command line?

The command line is your direct access to a computer. It's where you ask software to perform hardware actions that point-and-click graphical user interfaces (GUIs) simply can't ask.

Command lines are available on many operating systems—proprietary or open source. But it’s usually associated with Linux, because both command lines and open source software, together, give users unrestricted access to their computer.

**Use Ubuntu**

Ubuntu is a Linux-based operating system. It is designed for computers, smartphones, and network servers. The system is developed by a UK based company called Canonical Ltd. All the principles used to develop the Ubuntu software are based on the principles of Open Source software development.

## Features of Ubuntu

Following are some of the significant features of Ubuntu −

* The desktop version of Ubuntu supports all the normal software on Windows such as Firefox, Chrome, VLC, etc.
* It supports the office suite called **LibreOffice**.
* Ubuntu has an in-built email software called Thunderbird, which gives the user access to email such as Exchange, Gmail, Hotmail, etc.
* There are a host of free applications for users to view and edit photos.
* There are also applications to manage videos and it also allows the users to share videos.
* It is easy to find content on Ubuntu with the smart searching facility.
* The best feature is, it is a free operating system and is backed by a huge open source community.

**Data Warehouse-:**

A *data warehouse* is a type of [data management](https://www.oracle.com/in/database/what-is-data-management/) system that is designed to enable and support business intelligence (BI) activities, especially analytics. Data warehouses are solely intended to perform queries and analysis and often contain large amounts of historical data. The data within a data warehouse is usually derived from a wide range of sources such as application log files and transaction applications.

A data warehouse centralizes and consolidates large amounts of data from multiple sources. Its analytical capabilities allow organizations to derive valuable business insights from their data to improve decision-making. Over time, it builds a historical record that can be invaluable to data scientists and business analysts. Because of these capabilities, a data warehouse can be considered an organization’s “single source of truth.”

## **Benefits of a Data Warehouse**

Data warehouses offer the overarching and unique benefit of allowing organizations to analyze large amounts of variant data and extract significant value from it, as well as to keep a historical record.

Four unique characteristics (described by computer scientist William Inmon, who is considered the father of the data warehouse) allow data warehouses to deliver this overarching benefit. According to this definition, data warehouses are

* ****Subject-oriented.**** They can analyze data about a particular subject or functional area (such as sales).
* ****Integrated.**** Data warehouses create consistency among different data types from disparate sources.
* ****Nonvolatile.**** Once data is in a data warehouse, it’s stable and doesn’t change.
* ****Time-variant.**** Data warehouse analysis looks at change over time.

A well-designed data warehouse will perform queries very quickly, deliver high data throughput, and provide enough flexibility for end users to “slice and dice” or reduce the volume of data for closer examination to meet a variety of demands—whether at a high level or at a very fine, detailed level. The data warehouse serves as the functional foundation for middleware BI environments that provide end users with reports, dashboards, and other interfaces.

## **Data Warehouse Architecture**

The architecture of a data warehouse is determined by the organization’s specific needs. Common architectures include

* ****Simple.**** All data warehouses share a basic design in which metadata, summary data, and raw data are stored within the central repository of the warehouse. The repository is fed by data sources on one end and accessed by end users for analysis, reporting, and mining on the other end.
* ****Simple with a staging area.**** Operational data must be cleaned and processed before being put in the warehouse. Although this can be done programmatically, many data warehouses add a staging area for data before it enters the warehouse, to simplify data preparation.
* ****Hub and spoke.**** Adding data marts between the central repository and end users allows an organization to customize its data warehouse to serve various lines of business. When the data is ready for use, it is moved to the appropriate data mart.
* ****Sandboxes.**** Sandboxes are private, secure, safe areas that allow companies to quickly and informally explore new datasets or ways of analyzing data without having to conform to or comply with the formal rules and protocol of the data warehouse.

**Elasticsearch -:**

Elasticsearch is a real-time distributed and open source full-text search and analytics engine. It is used in Single Page Application (SPA) projects. Elasticsearch is an open source developed in Java and used by many big organizations around the world. It is licensed under the Apache license version 2.0.

Elasticsearch is an Apache Lucene-based search server. It was developed by Shay Banon and published in 2010. It is now maintained by Elasticsearch BV. Its latest version is 7.0.0.

Elasticsearch is a real-time distributed and open source full-text search and analytics engine. It is accessible from RESTful web service interface and uses schema less JSON (JavaScript Object Notation) documents to store data. It is built on Java programming language and hence Elasticsearch can run on different platforms. It enables users to explore very large amount of data at very high speed.

The general features of Elasticsearch are as follows −

* Elasticsearch is scalable up to petabytes of structured and unstructured data.
* Elasticsearch can be used as a replacement of document stores like MongoDB and RavenDB.
* Elasticsearch uses denormalization to improve the search performance.
* Elasticsearch is one of the popular enterprise search engines, and is currently being used by many big organizations like Wikipedia, The Guardian, StackOverflow, GitHub etc.
* Elasticsearch is an open source and available under the Apache license version 2.0.

## Key Concepts

The key concepts of Elasticsearch are as follows −

### Node

It refers to a single running instance of Elasticsearch. Single physical and virtual server accommodates multiple nodes depending upon the capabilities of their physical resources like RAM, storage and processing power.

### Cluster

It is a collection of one or more nodes. Cluster provides collective indexing and search capabilities across all the nodes for entire data.

### Index

It is a collection of different type of documents and their properties. Index also uses the concept of shards to improve the performance. For example, a set of document contains data of a social networking application.

### Document

It is a collection of fields in a specific manner defined in JSON format. Every document belongs to a type and resides inside an index. Every document is associated with a unique identifier called the UID.

### Shard

Indexes are horizontally subdivided into shards. This means each shard contains all the properties of document but contains less number of JSON objects than index. The horizontal separation makes shard an independent node, which can be store in any node. Primary shard is the original horizontal part of an index and then these primary shards are replicated into replica shards.

### Replicas

Elasticsearch allows a user to create replicas of their indexes and shards. Replication not only helps in increasing the availability of data in case of failure, but also improves the performance of searching by carrying out a parallel search operation in these replicas.

## Advantages

* Elasticsearch is developed on Java, which makes it compatible on almost every platform.
* Elasticsearch is real time, in other words after one second the added document is searchable in this engine
* Elasticsearch is distributed, which makes it easy to scale and integrate in any big organization.
* Creating full backups are easy by using the concept of gateway, which is present in Elasticsearch.
* Handling multi-tenancy is very easy in Elasticsearch when compared to Apache Solr.
* Elasticsearch uses JSON objects as responses, which makes it possible to invoke the Elasticsearch server with a large number of different programming languages.
* Elasticsearch supports almost every document type except those that do not support text rendering.

## Disadvantages

* Elasticsearch does not have multi-language support in terms of handling request and response data (only possible in JSON) unlike in Apache Solr, where it is possible in CSV, XML and JSON formats.
* Occasionally, Elasticsearch has a problem of Split brain situations.

## Comparison between Elasticsearch and RDBMS

In Elasticsearch, index is similar to tables in RDBMS (Relation Database Management System). Every table is a collection of rows just as every index is a collection of documents in Elasticsearch.

**Scala -:**

Scala programming is a general-purpose computer language that supports both object-oriented and functional styles of programming on a larger scale. Scala is a strong static type of programming language and is influenced by the Java programming language. One of the best similarities of Scala and Java is that you can code Scala just the same

way that you code Java. It is also possible to use a lot of Java libraries within Scala along with many of its third-party libraries.

Scala programming is a general-purpose computer language that supports both object-oriented and functional styles of programming on a larger scale. Scala is a strong static type of programming language and is influenced by the Java programming language. One of the best similarities of Scala and Java is that you can code Scala just the same way that you code Java. It is also possible to use a lot of Java libraries within Scala along with many of its third-party libraries.

**Advantage of Scala**

* The code is more concise, readable, and error-free.
* It is easy to write, compile, debug, and run the program in Scala, when compared to many other programming languages.
* Functional programming lets you approach the same problem in a different angle.
* It deploys concurrency which helps in parallelizing tasks.
* A lot of third-party libraries can be utilized for specific tasks. They can be added in the form of language constructs.

****Some of the ways in which Scala language is deployed in real-world scenarios are as follows:****

* Writing web applications
* Applications working with streaming data
* Concurrency and distributed applications
* Parallel batch processing
* Analyzing data with [Apache Spark](https://intellipaat.com/blog/what-is-apache-spark/" \t "https://intellipaat.com/blog/what-is-scala/_blank)

## ****Scope of Scala****

Since its inception, Scala has witnessed tremendous growth in its demand, which is sure to increase in the near future. Some of the statistics that support this hypothesis are mentioned below:

* Scala’s popularity has been constantly growing for the last two years. Almost 48 percent of the respondents would like to work with Scala this year**–**JAXenter survey
* Scala ranks high in salary surveys, coming up among the top 10 worldwide and sharing the first place in the US – Stack Overflow’s report

**Tool Short listing**

**PyCharm Community edition 2020.1.2**

PyCharm is a dedicated Python Integrated Development Environment (IDE) providing a wide range of essential tools for Python developers, tightly integrated to create a convenient environment for productive Python, web, and data science development.

PyCharm is available in three editions:

1. *Community* (free and [open-sourced](https://github.com/JetBrains/intellij-community/blob/master/LICENSE.txt)): for smart and intelligent Python development, including code assistance, refactorings, visual debugging, and version control integration.
2. *Professional* ([paid](https://www.jetbrains.com/pycharm/buy/" \l "commercial?billing=yearly)) : for professional Python, web, and data science development, including code assistance, refactorings, visual debugging, version control integration, remote configurations, deployment, support for popular web frameworks, such as Django and Flask, database support, scientific tools (including Jupyter notebook support), big data tools.
3. *Edu* (free and [open-sourced](https://github.com/JetBrains/intellij-community/blob/master/LICENSE.txt)): for learning programming languages and related technologies with integrated educational tools.

**VirtualBox**

VirtualBox is [open-source](https://www.computerhope.com/jargon/o/opensour.htm) software for virtualizing the [x86](https://www.computerhope.com/jargon/x/x86.htm) computing [architecture](https://www.computerhope.com/jargon/a/architec.htm). It acts as a [hypervisor](https://www.computerhope.com/jargon/h/hypervisor.htm), creating a [VM](https://www.computerhope.com/jargon/v/virtmach.htm) (virtual machine) where the user can run another [OS](https://www.computerhope.com/jargon/o/os.htm) (operating system).

The operating system where VirtualBox runs is called the "host" OS. The operating system running in the VM is called the "guest" OS. VirtualBox supports [Windows](https://www.computerhope.com/jargon/w/windows.htm), [Linux](https://www.computerhope.com/jargon/l/linux.htm), or [macOS](https://www.computerhope.com/jargon/m/macos.htm) as its host OS.

When configuring a virtual machine, the user can specify how many [CPU](https://www.computerhope.com/jargon/c/cpu.htm) [cores](https://www.computerhope.com/jargon/c/core.htm), and how much [RAM](https://www.computerhope.com/jargon/r/ram.htm) and [disk space](https://www.computerhope.com/jargon/d/diskcapa.htm) should be devoted to the VM. When the VM is running, it can be "paused." System [execution](https://www.computerhope.com/jargon/e/execute.htm) is frozen at that moment in time, and the user can resume using it later.

## History

VirtualBox was originally developed by Innotek GmbH, and released on January 17, [2007](https://www.computerhope.com/history/2007.htm) as an open-source software package. The company was later purchased by [Sun Microsystems](https://www.computerhope.com/comp/sun.htm).

On January 27, 2010, [Oracle Corporation](https://www.computerhope.com/comp/oracle.htm) purchased Sun, and took over development of VirtualBox.

## Supported guest operating systems

Guest operating systems supported by VirtualBox include:

1. [Windows 10](https://www.computerhope.com/jargon/w/windows-10.htm), [8](https://www.computerhope.com/jargon/w/windows8.htm), [7](https://www.computerhope.com/jargon/w/windows7.htm), [XP](https://www.computerhope.com/jargon/w/winxp.htm), [Vista](https://www.computerhope.com/jargon/v/vista.htm), [2000](https://www.computerhope.com/jargon/w/win2000.htm), [NT](https://www.computerhope.com/jargon/w/winnt.htm), and [98](https://www.computerhope.com/jargon/w/win98.htm).
2. Linux distributions based on [Linux kernel](https://www.computerhope.com/jargon/l/linux-kernel.htm) 2.4 and newer, including [Ubuntu](https://www.computerhope.com/jargon/u/ubuntu.htm), [Debian](https://www.computerhope.com/jargon/d/debian.htm), [OpenSUSE](https://www.computerhope.com/jargon/s/suse-linux.htm), [Mandriva](https://www.computerhope.com/jargon/m/mandriva-linux.htm)/Mandrake, [Fedora](https://www.computerhope.com/jargon/f/fedora-linux.htm), RHEL, and [Arch Linux](https://www.computerhope.com/jargon/a/arch-linux.htm).
3. [Solaris](https://www.computerhope.com/jargon/s/solaris.htm) and OpenSolaris.
4. macOS X Server Leopard and Snow Leopard.
5. [OpenBSD](https://www.computerhope.com/jargon/o/openbsd.htm) and [FreeBSD](https://www.computerhope.com/jargon/b/bsd.htm" \l "freebsd).
6. [MS-DOS](https://www.computerhope.com/jargon/m/msdos.htm).
7. [OS/2](https://www.computerhope.com/jargon/o/os2.htm).
8. [QNX](https://www.computerhope.com/jargon/q/qnx.htm).
9. [BeOS](https://www.computerhope.com/jargon/b/beos.htm) R5.
10. Haiku.
11. [ReactOS](https://www.computerhope.com/jargon/r/reactos.htm).

**Putty**

PuTTY is a free implementation of  SSH (and telnet) for PCs running Microsoft Windows (it also includes an xterm terminal emulator). You will find PuTTY useful if you want to access an account on a Unix or other multi-user system from a PC (for example your own or one in an internet cafe). In the case of the University of Sussex, this would include users of the central ITS Unix system and of the Unix systems provided by some other schools and units. The central ITS Unix system supports the use of SSH clients. Users of other systems should ask their system administrator if SSH is supported.

PuTTY is an alternative to telnet clients. Its primary advantage is that SSH provides a secure, encrypted connection to the remote system. It's also small and self-contained and can be carried around on a floppy disk. This makes it ideal for accessing Sussex systems securely from other locations on the public Internet.

**WinSCP**

WinSCP is an open source free [SFTP client](https://winscp.net/eng/docs/free_sftp_client_for_windows), [FTP client](https://winscp.net/eng/docs/free_ftp_client_for_windows), [WebDAV](https://winscp.net/eng/docs/webdav) client, [S3](https://winscp.net/eng/docs/s3) client and [SCP](https://winscp.net/eng/docs/scp) client for Windows. Its main function is file transfer between a local and a remote computer. Beyond this, WinSCP offers scripting and basic file manager functionality.

WinSCP (Windows Secure Copy) is an open source SecureFTP client for Windows. It allows secure file transfers between the client's local computer and the remote server. The WinSCP software uses cryptographical methods, integrated into SSH to protect your login details and private information.

**MySQL Workbench**

MySQL Workbench is a unified visual tool for database architects, developers, and DBAs. MySQL Workbench provides data modeling, SQL development, and comprehensive administration tools for server configuration, user administration, backup, and much more. MySQL Workbench is available on Windows, Linux and Mac OS

## Develop

MySQL Workbench delivers visual tools for creating, executing, and optimizing SQL queries. The SQL Editor provides color syntax highlighting, auto-complete, reuse of SQL snippets, and execution history of SQL. The Database Connections Panel enables developers to easily manage standard database connections, including MySQL Fabric. The Object Browser provides instant access to database schema and objects.

## Administer

MySQL Workbench provides a visual console to easily administer MySQL environments and gain better visibility into databases. Developers and DBAs can use the visual tools for configuring servers, administering users, performing backup and recovery, inspecting audit data, and viewing database health.

**Intellij IDEA**

**I**ntelliJ IDEA is an *Integrated Development Environment (IDE)* for JVM languages designed to maximize developer productivity. It does the routine and repetitive tasks for you by providing clever code completion, static code analysis, and refactorings, and lets you focus on the bright side of software development, making it not only productive but also an enjoyable experience.

## Multi-platform

IntelliJ IDEA is a cross-platform IDE that provides consistent experience on Windows, macOS, and Linux.

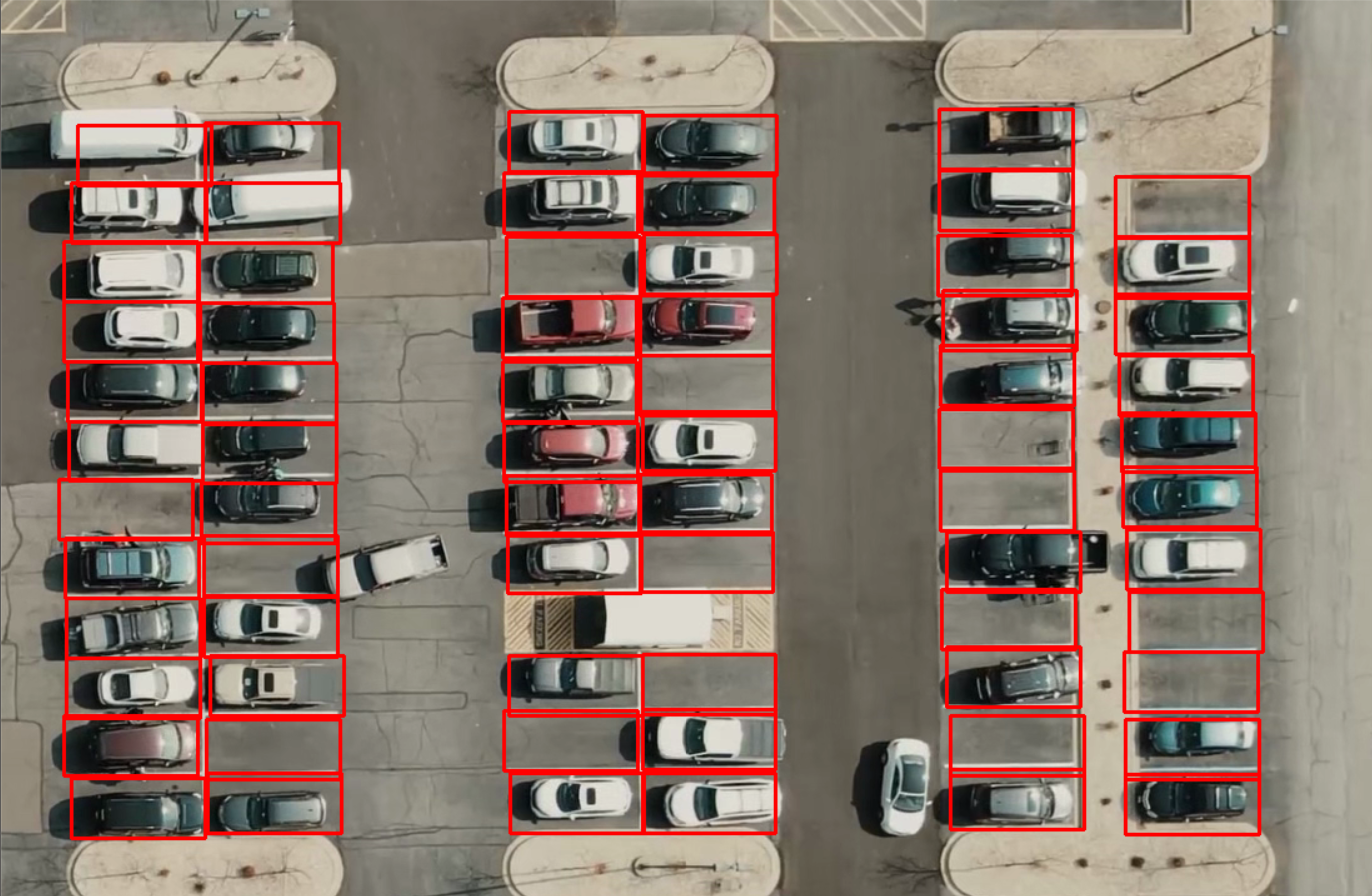
* See [Install IntelliJ IDEA](https://www.jetbrains.com/help/idea/installation-guide.html) for OS-specific instructions.
* See [IntelliJ IDEA keyboard shortcuts](https://www.jetbrains.com/help/idea/mastering-keyboard-shortcuts.html) for instructions on how to choose the right keymap for your operating system, and learn the most useful shortcuts.

**Chapter 4**

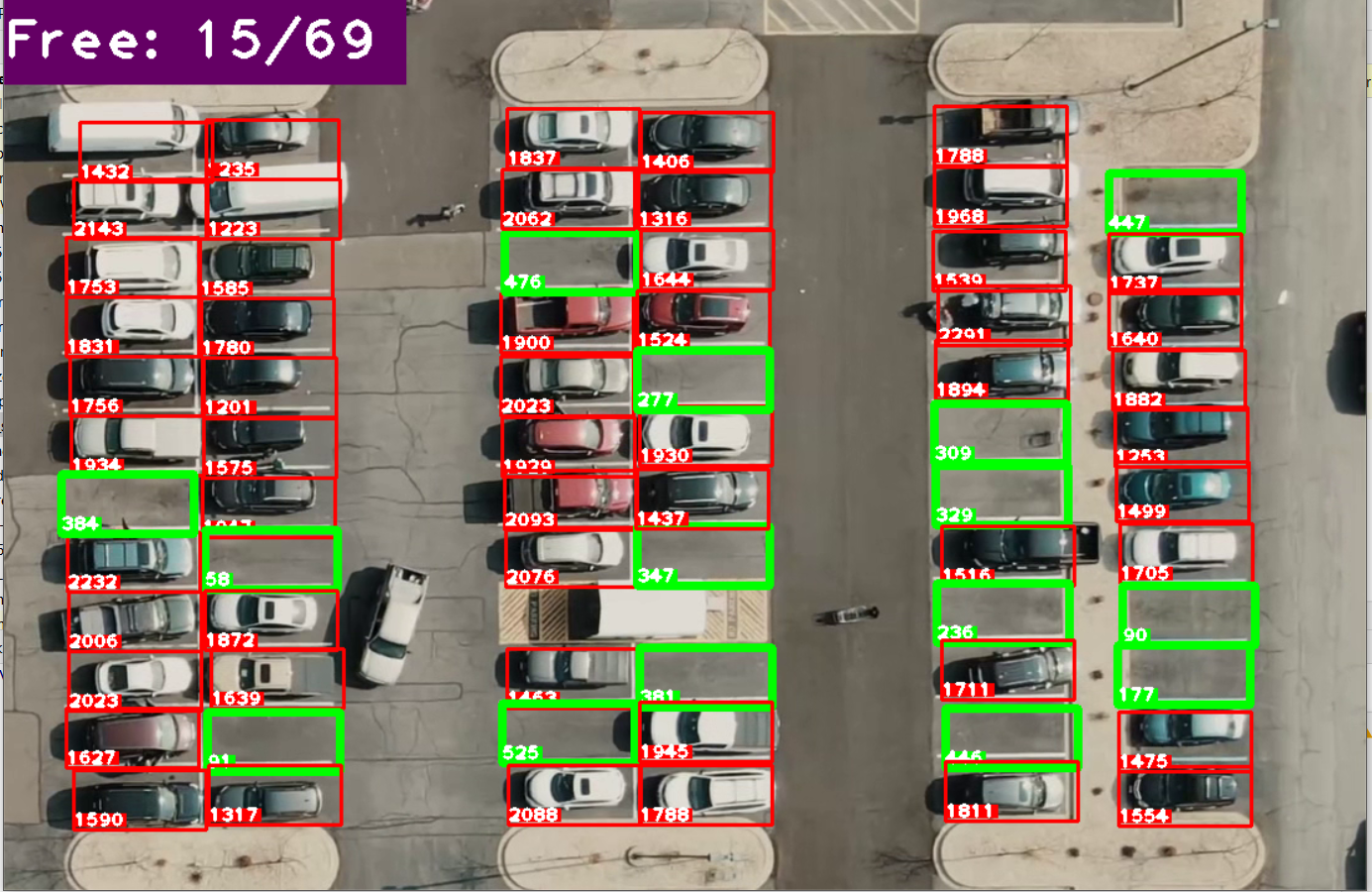
**Project methodology**



This page represent how the front end of the application looks build by using python tkinter.

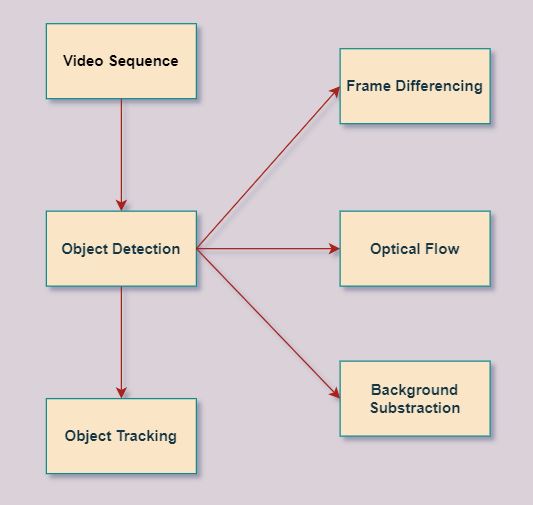


**Red\_Rectangle-:** The red rectangle in this picture depicts the area where a car can park and this can be selected or deselected by just clicking your mouse left click or by right click.

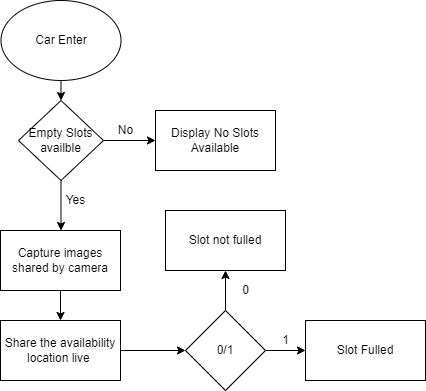


**Green\_Rectangle-:** The green rectangle depicts the area where a car parking is available.

**Free: 15/69-:** The number representing 15 out of 69 places are free to park your car.



Above picture depicts a simple diagram of the discovery and tracing. In paper, algorithms based on SSDs and MobileNets are used to locate and track python location. Object discovery involves finding the object of interest of an object in a particular part of an image. Various methods of frame separation, light flow, background removal. This is a way to find something moving with the support of a camera.



when a car enters the first task is to check whether there is any slot available If “No” it simply return No slots available, If condition is “Yes” than it capture the image frame wise and pass through the certain algorithm which detect whether the slot is filled or empty. If found filled than represented by red rectangle and If green it means they are vacant and now there you can park your car. If any filled slot become available than the application automatically do it’s task and show vacant on screen through a green rectangle(0).

**Chapter 5**

**Project Challenges**

2- Months Advanced Technical Training was conducted

Modules of Training:

1. Prerequisites technology(Python, Linux and hadoop)
2. Big Data(Hadoop) module
3. Pyspark module
4. Business Training(Focused on communication skills)

With Test(KBA) Conducted for each module and case study base assignment given to all interns in which some task is given which we have to perform by using all those skills which we have go through during the internship periods. And, the 2nd thing is learn and become a better communicator through twice a week business training provided by well skilled senior employee of cognizant.

**Benefit to the company**

The internship is completely self paced so this will not directly benefited to company indeed the main focus of this internship is to get hands on knowledge about the tech which we are going to work. So, the benefit to company is they get skilled employee.

**Chapter 6**

* How the objectives achieved?

So, to achieve the objective we need to complete the prerequisites courses in a week and than we move to the next step which is for main course and there we have around 2 months to complete all the courses by daily investing and progressing of 4 hrs. In mid there are some udemy based test where we have to maintain the rack of 70% and above and after everyone complete there main courses we are passed through a knowledge based assessment with the same criteria of 70% and case studies where we have given some problems which we need to find it’s solutions.

• What skills (scientific and professional) were learned during the internship?

During the period of internship I learned technical skill like Python, Hadoop environment(HBase, Hive, Map reduce, Hbase, kafka ….), Linux, Pyspark and Scala. Add on to no technical but professional skill learn how to communicate with colleagues in a connect and also get some professional etiquette.

• Results/observations/work experiences get in the internship company.

The internship is completely self faced so learning by itself and getting the access to premium Udemy courses for free makes my learning better and, getting a domain of most demand technology like big data makes me future ready.

• What challenges did you experience during the internship?

Initially, I faces a management challenge like for first few month I’m not allocated to any domain after that they put me in digital marketing domain and again switch me to Big data & Pyspark. With all these mismanagement It take too long to get any reply from mentor. Than challenges in terms of technology and platform big data required high computation power which my system is not fully compatible and by company side get the access to platform after a month when my technical training started and, more worst is the platform isn’t work, while raising a question they replied bid data lab is facing some technical issue.

**References -:**

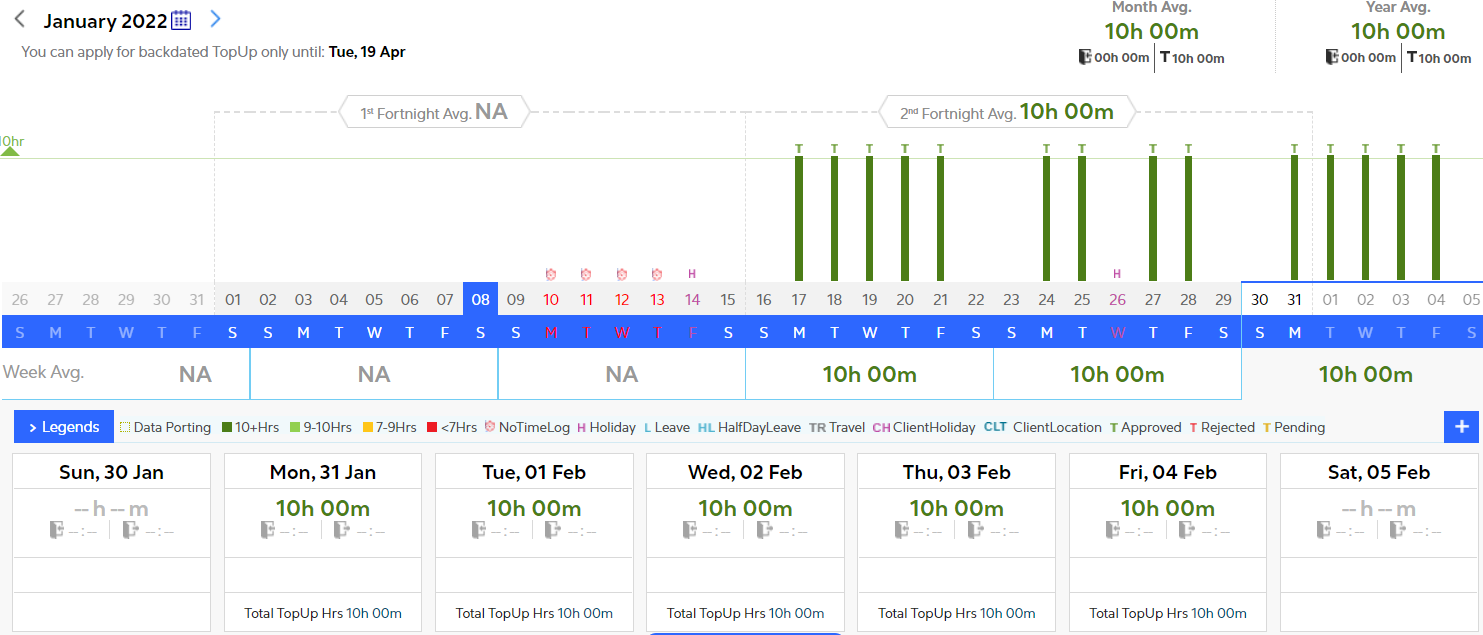
https://cognizantlearning.sumtotal.host/rcore/c/dash/home?relyingParty=ELIXHRUI&nodeKey=dashboard\_1\_Dashboard&nodeUrl=%2Fdash%2Fhome

https://www.tutorialspoint.com/elasticsearch/elasticsearch\_basic\_concepts.htm

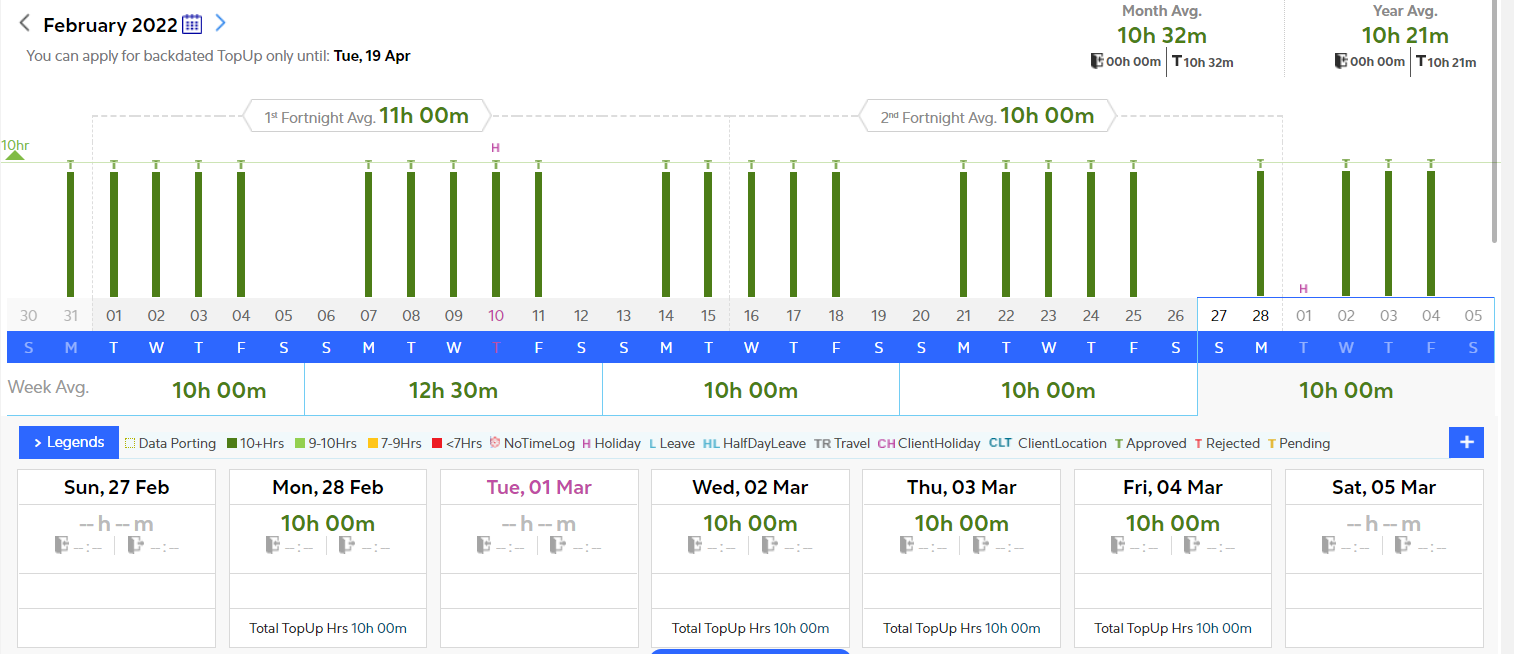
**Annexure –I Attendance Report - :**

True time regular attendance platform

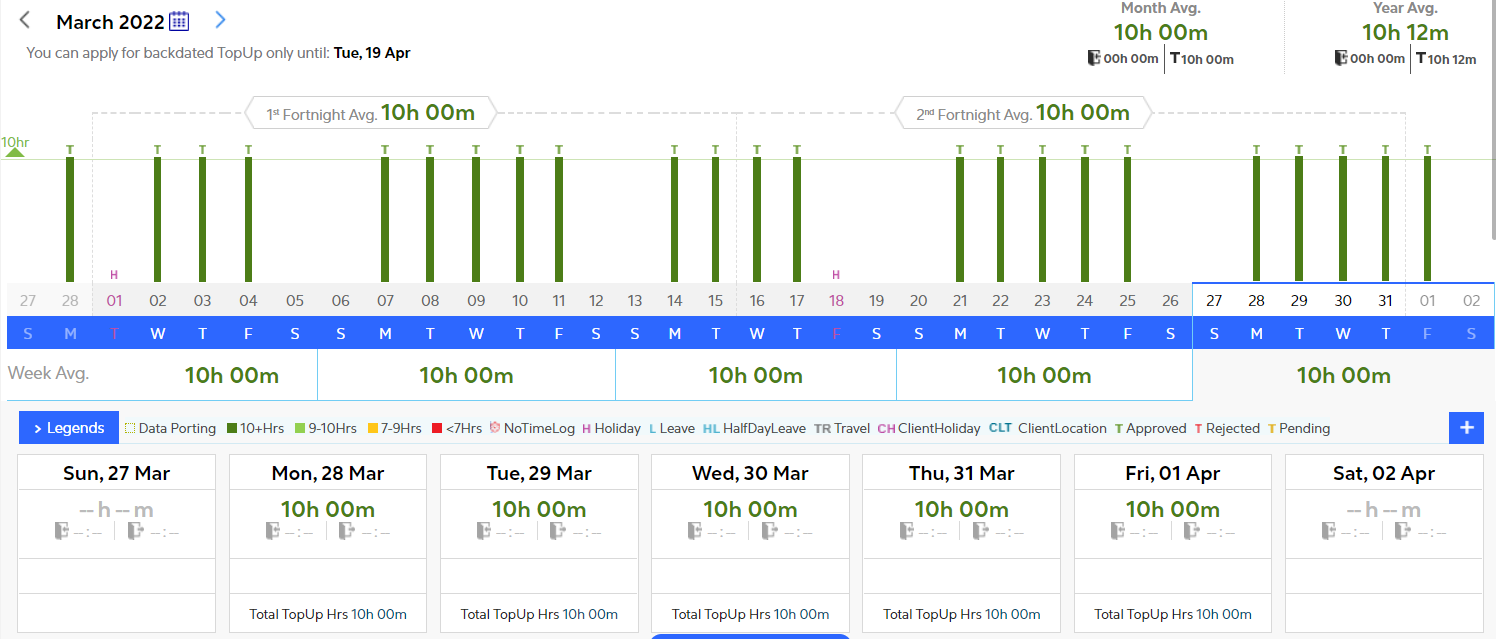
**January**



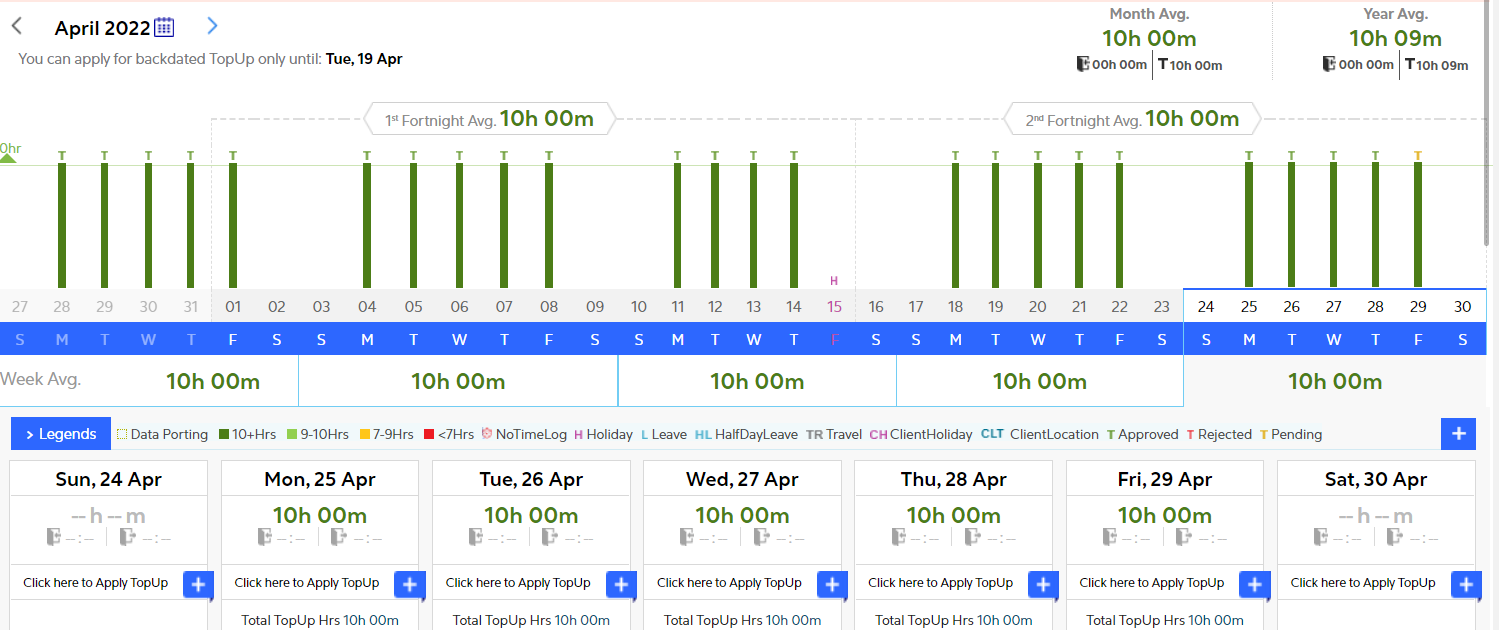
February -:



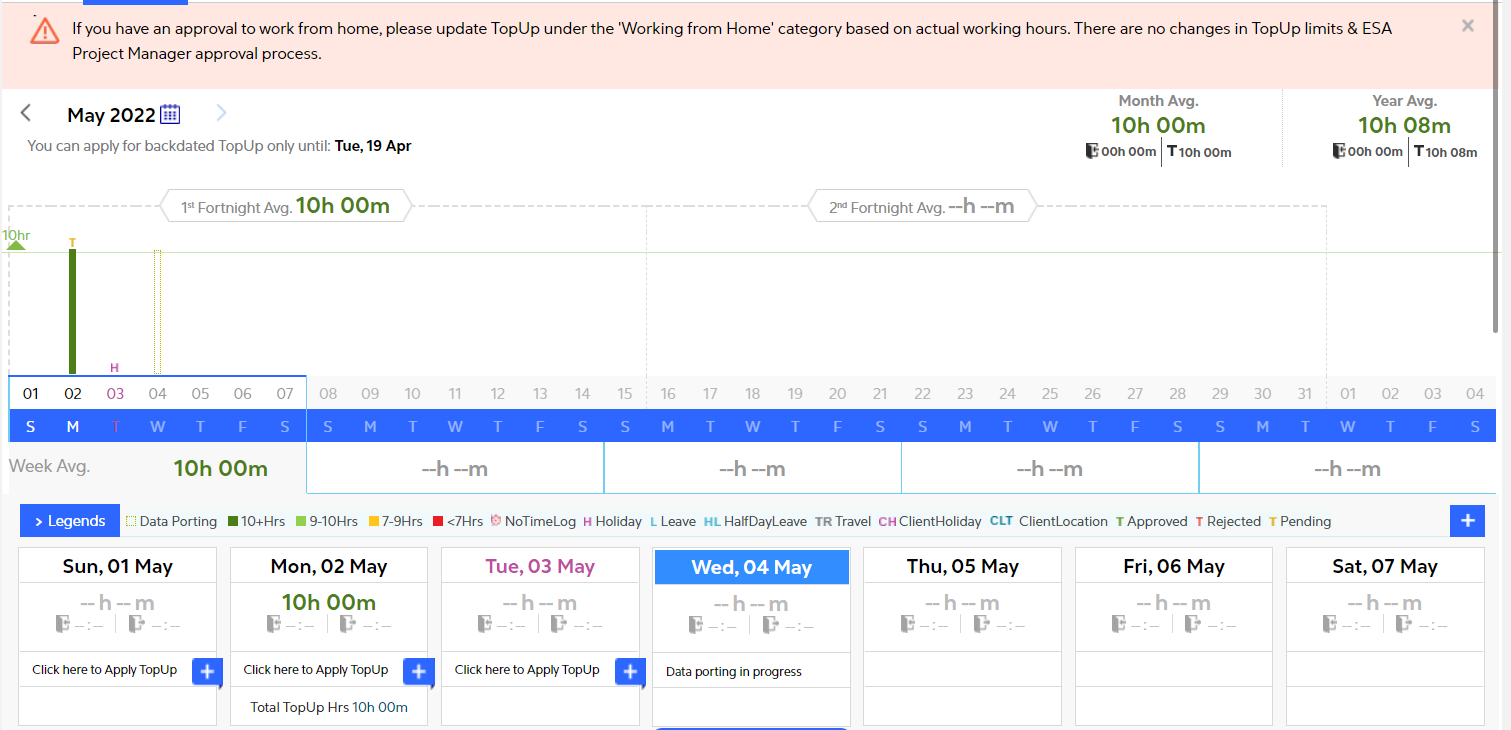
**March-:**



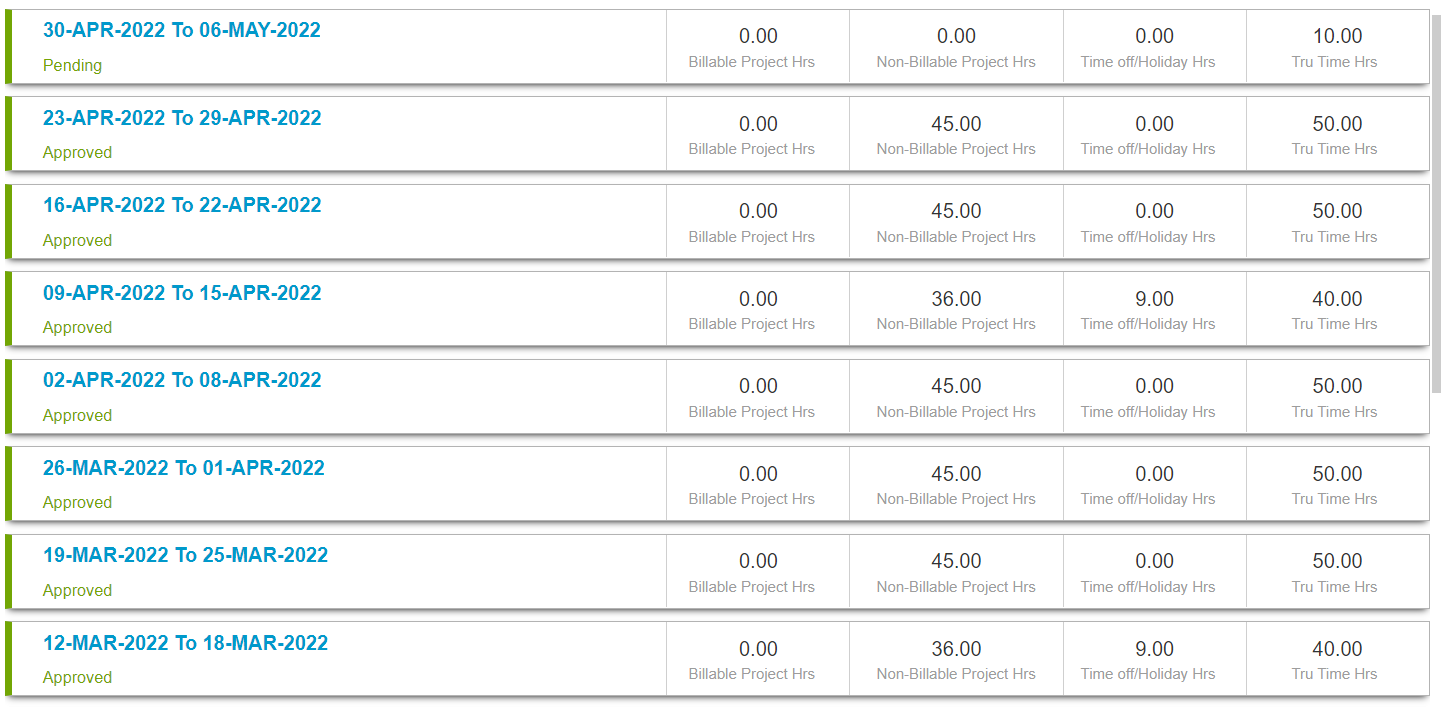
**April-:**

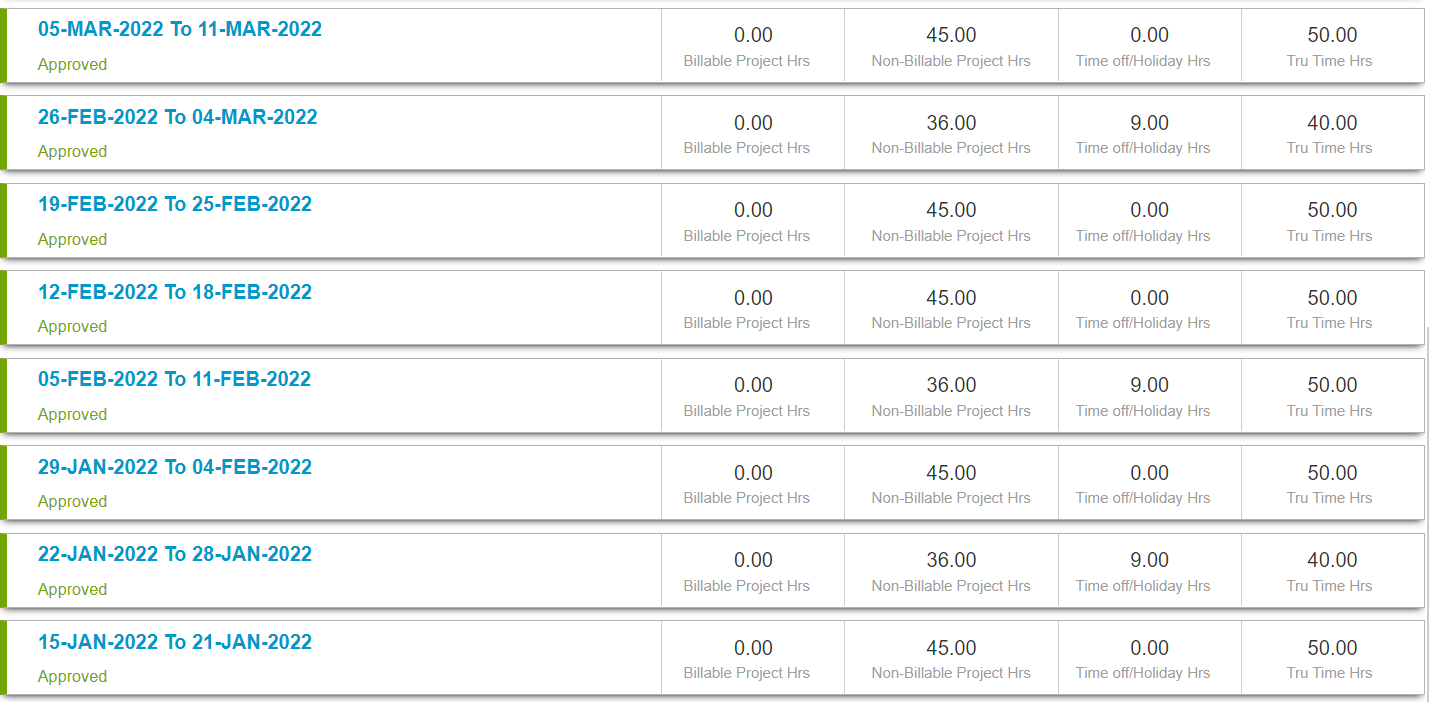


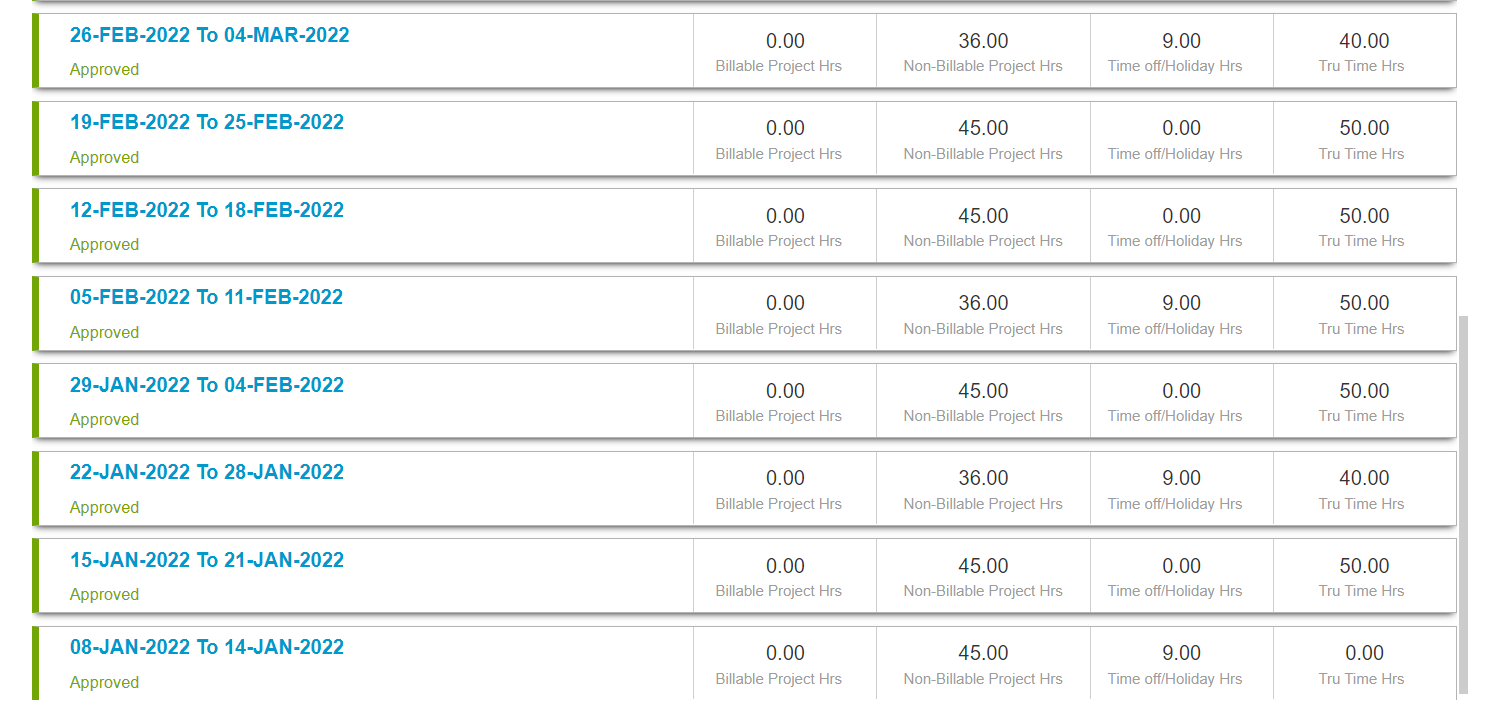
**May -:**



Time sheet weekly attendance platform

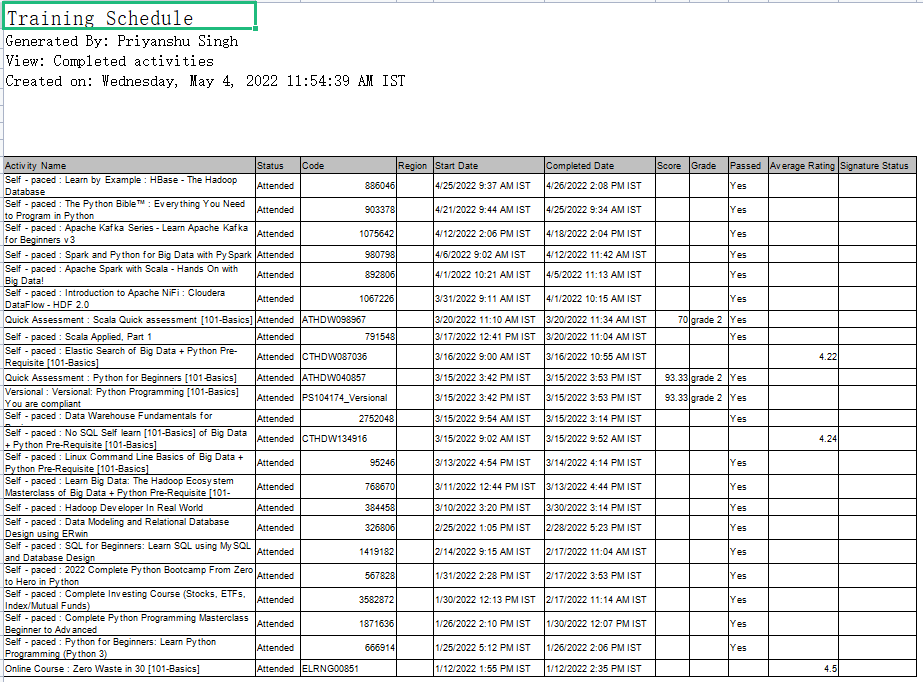






**Annexure –II**

Weekly overview of internship activities



**Annexure –III Salary Slips -:**

