# PHASE - II

## *CREDIT CARD FRAUD DETECTION:*

Credit card fraud detection is a set of methods and techniques designed to block fraudulent purchases, both online and in-store. This is done by ensuring that the myriads of plastic cards in use worldwide are a gold mine for criminals by2027, financial service providers are expected to take a $40 billion hit globally in credit card losses, a significant increase compared to $27.85 bn  in 2018.

This growth in losses is partially caused by the rise of electronic transactions. Just imagine that today the average American has more than three credit cards, which amounts to 1.5 billion cards in the US alone. While the number of plastic cards globally numbers an impressive 22.11 billion.

Another reason is that fraudulent methods are getting more sophisticated and thus harder to spot by traditional fraud detection software.

### *Targets of Credit Card Fraud Detection*

credit card fraud is *"the unauthorized use of a credit or debit card, or similar payment tool to fraudulently obtain money or property."* All players involved in the card-based payment process can potentially fall victim to scammers, including:

- cardholders,
- online merchants,
- payment gateway providers,
- payment processing companies,
- credit card payment systems,
- card issuers (issuing banks), and
- acquirers (acquiring banks).

### *Machine Learning Techniques*

- **Supervised learning** means that a model learns from previous examples and is trained on labeled data. In other words, the dataset has tags that tell the model which patterns are related to fraud and which represent normal behavior.
- *"Banks and payment systems typically accumulate tons of data on different fraudulent schemes that can be used to train a model,"* Alexander Konduforov says. *"Such models are constantly updated and improved to produce accurate results. But unfortunately, they fail to spot new fraud schemes if faced with them."* That's when unsupervised learning comes into the picture.

- **Unsupervised learning** is also called anomaly detection as it automatically captures unusual patterns. In this case, training datasets come without any labels or instructions. This approach lags behind supervised learning in terms of accuracy. But it is unrivaled when a business needs to find hidden fraud patterns and useful insights.

  ***TYPES:***

- **Stealing cards** – This is a standard method of committing a credit card fraud transaction where the fraudster gets possession of the card and can directly misuse it.
- **Hacking email accounts** – Hackers devise programs or software to hack into email accounts because banks send emails to customers with account information.
- **Obtain details by calling the cardholder** – Sometimes fraudsters call customers, lure them with gifts or facilities, and convince them to reveal credit card information.
- **Obtain access to the bank account**– Sometimes, it is possible to obtain access to the bank account if the fraudster is a person the account holder knows and trusts.
- **Phishing through messages having fraudulent links/codes** – Cardholders may get mobile messages or emails containing unknown links, which, if clicked, will directly **debit** money from bank accounts, which is a standard method of committing credit card fraud online.
- **Skimming and cloning**– Many automated teller machines (ATM) or card swiping machines are fitted with devices that copy the card information as soon as it is swiped, later used to design a fake card.
- **Donation request** – Cardholders often get calls asking for financial help to save people in need. It is necessary to verify the caller before transferring money in such cases.
- **Scarring cardholders with arrest calls** – Scammers try to scare cardholders with pending payments with arrest threats and ask for extra payment as a penalty, which is a false claim.
- **Attracting cardholders with offers and interest reductions** – Cardholders get calls offering them reduced **interest** on pending payments if they transfer some amount to an account number.
- **Claiming overpayment** – Sometimes, fraud calls claim that cardholders overpaid for some purchase, for which they will get a refund provided they give their card details.

## *ABOUT DATASETS*:

Kaggle is one of the largest communities of data scientists and machine learning practitioners in the world, and its platform hosts thousands of datasets covering a wide range of topics and industries. With so many options to choose from, it can be difficult to know where to start or what datasets are worth exploring. That's where this dataset comes in. By scraping information about the top 10,000 datasets on Kaggle, we have created a single source of truth for the most popular and useful datasets on the platform. This dataset is not just a list of names and numbers, but a valuable tool for data enthusiasts and professionals alike, providing insights into the latest trends and techniques in data science and machine learning.

### *Column Descriptions:*

- **Dataset_name** - Name of the dataset
- **Author_name** - Name of the author
- **Author_id** - Kaggle id of the author
- **No_of_files** - Number of files the author has uploaded
- ***size*** - Size of all the files
- **Type_of_file** - Type of the files such as csv, json etc.
- **Upvoste** - Total upvotes of the dataset
- **Medals -** Medal of the dataset
- **Usability** - Usability of the dataset
- **Date -** Date in which the dataset is uploaded.
- **Day** - Day in which the dataset is uploaded.
- **Time -** Time in which the dataset is uploaded.
- **Dataset_link** - Kaggle link of the dataset

## WHAT IS PANDA IN PYTHON?

Pandas is an open-source Python library developed by Wes McKinney in 2008. It is used in data science, data analysis, and other machine-learning activities. It is very fast and provides many tools for effectively handling large amounts of data. It is built on the Numpy library.

### *Why using pandas in Python?*

Pandas strengthens Python by giving the popular programming language the capability to work with spreadsheet-like data enabling fast loading, aligning, manipulating, and merging, in addition to other key functions. PANDAS is short for Pediatric Autoimmune Neuropsychiatric Disorders Associated with Streptococcal Infections. A child may be diagnosed with PANDAS when: Obsessive-compulsive disorder (OCD), tic disorder, or both suddenly appear following a streptococcal (strep) infection, such as strep throat or scarlet fever.

### *What is Pandas and its types?*

Pandas works with many different types of data sets such as comma-separated values (CSV) files, Excel files, extensible markup language (XML) files, JavaScript object notation (JSON) files and relational database tables. Data read from these sources are returned as Pandas data types known as DataFrame and Series. Pandas a Python library? Pandas is a Python library for data analysis. Started by Wes McKinney in 2008 out of a need for a powerful and flexible quantitative analysis tool, pandas has grown into one of the most popular Python libraries.

## ***Who uses Python pandas?***

**Data Scientists:**
The pandas package is the most important tool at the disposal of Data Scientists and Analysts working in Python today. The powerful machine learning and glamorous visualization tools may get all the attention, but pandas is the backbone of most data projects. How pandas are used in Python? Pandas is a Python library used for working with data sets. It has functions for analyzing, cleaning, exploring, and manipulating data. The name "Pandas" has a reference to both "Panel Data", and "Python Data Analysis" and was created by Wes McKinney in 2008.

Method: Using pip

It's a package installation tool that simplifies the installation of Python modules and frameworks.Pip will be installed with Python by default if you have a later version of Python available (greater than Python 3.5.x). If you're using an earlier version of Python, you'll need to install pipbefore you can install Pandas. The simplest method to accomplish this is to update to the most recent version of Python, which can be found at this link.

**Step 1: Launch Command Prompt**

To open the start menu, use the Windows key on your keyboard or click the Start button. For example, when you type "cmd" the Command Prompt app should display in the start menu, and once you can view the command prompt app, launch the app. Alternatively, you may hit the Windows key + r to bring up the "RUN" box, where you can input “cmd” and then press enter. Itwill also launch the Command prompt.

**Step 2: Enter the command**

After you open the command prompt, the following step is to enter the needed command to begin the pip installation. For example, enter the command shown below. This will start the pip installation. After downloading the necessary files, Pandas will be set to operate onyour computer. You can employ Pandas in your Python projects once the installation has been completed.

## ***REST OF EXPLANATION:***

RESTful API is an architectural style for an application program interface (API) that uses HTTP requests to access and use data. That data can be used to GET, PUT, POST and DELETE data types, which refers to the reading, updating, creating and deleting of operations concerning.

A RESTful API uses commands to obtain resources. The state of a resource at any given timestamp is called a resource representation. A RESTful API uses existing HTTP methodologies defined by the RFC 2616 protocol, such as:

- GET to retrieve a resource;
- PUT to change the state of or update a resource, which can be an object, file or block;
- POST to create that resource;
- DELETE to remove it

**Data formats the REST API supports include:**

- application/json
- application/xml
- application/x-wbe+xml
- application/x-www-form-urlencoded
- multipart/form-data

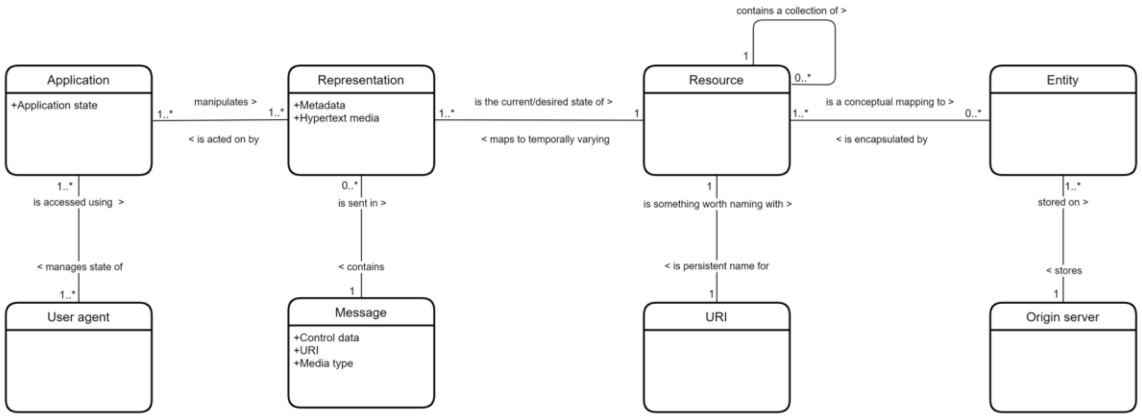
### ***USES:***

Because the calls are stateless, REST is useful in cloud applications. Stateless components can be freely redeployed if something fails, and they can scale to accommodate load changes. This is because any request can be directed to any instance of a component; there can be nothing saved that has to be remembered by the next transaction. That makes REST preferable for web use. The RESTful model is also helpful in cloud services because binding to a service through an API is a matter of controlling how the URL is decoded. Cloud computing and microservices are almost certain to make RESTful API design the rule.

### ARCHITECTURAL PROPERTIES:

The REST architectural style is designed for network-based applications, specifically client-server applications. But more than that, it is designed for Internet-scale usage, so the coupling between the user agent (client) and the origin server must be as loose as possible to facilitate large-scale adoption.

The strong decoupling of client and server together with the text-based transfer of information using a uniform addressing protocol provided the basis for meeting the requirements of the Web: robustness (anarchic scalability), independent deployment of components, large-grain data transfer, and a low entry-barrier for content readers,



An entity-relationship model of the concepts expressed in the REST architectural style

The constraints of the REST architectural style affect the following architectural properties:

- Performance in component interactions, which can be the dominant factor in user-perceived performance and network efficiency.
- Scalability allowing the support of large numbers of components and interactions among components.
- Simplicity of a uniform interface.
- Modifiability of components to meet changing needs (even while the application is running.
- Visibility of communication between components by service agents.
- Portability of components by moving program code with the data.
- Reliability in the resistance to failure at the system level in the presence of failures within components, connectors, or data.

## *What is in a dataset?*

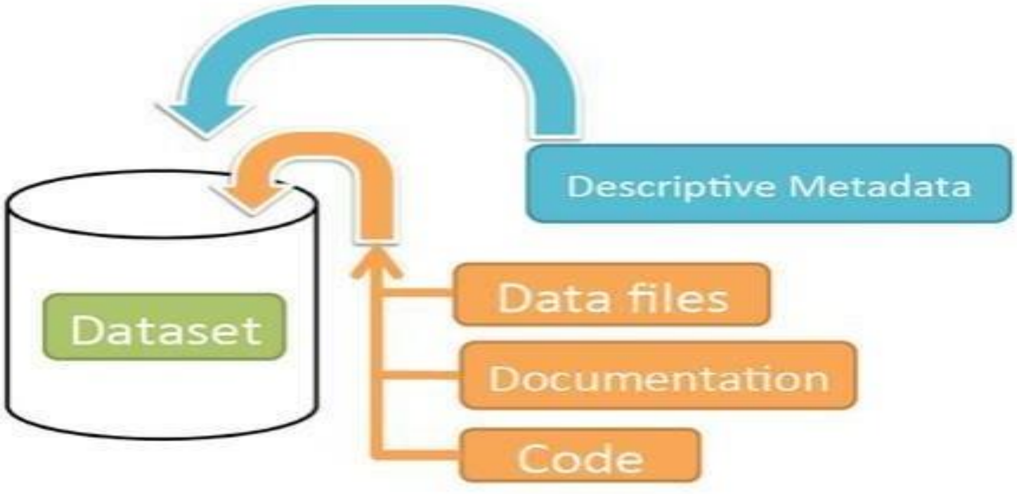
A data set is an ordered collection of data. As we know, a collection of information obtained through observations, measurements, study, or analysis is referred to as data. It could include information such as facts, numbers, figures, names, or even basic descriptions of objects.

## *What is dataset with example?*

A data set is a collection of numbers or values that relate to a particular subject. For example, the test scores of each student in a particular class is a data set. The number of fish eaten by each dolphin at an aquarium is a data set.

## *Why do we use dataset?*

Schematic Diagram of a **Dataset** in Dataverse 4.0



Container for your data, documentation, and code.

You can use datasets to conduct market research, analyze competitors, compare prices, identify and study trends, or train machine learning models. These are just a few examples, and datasets are useful in various areas and situations.

***Other types of data sets***:

•Numerical data, also known as quantitative data, is expressed in numbers instead of in what we know as natural language.

•Bivariate data sets contain only two variables.

•Multivariate data sets contain at least three variables that are somehow related.

## *What is data set in Python?*

• A Dataset is the basic data container in PyMVPA. It serves as the primary form of data storage, but also as a common container for results returned by most algorithms. In this tutorial part we will take a look at what a dataset consists of, and how it work

**Construct your dataset (and before doing data transformation), you should:**

1. Collect the raw data.
2. Identify feature and label sources.
3. Select a sampling strategy.
4. Split the data.

***What are features of a dataset?***

5. Features defines the internal structure of a dataset. It is used to specify the underlying serialization format. What's more interesting to you though is that Features contains high-level information about everything from the column names and types, to the ClassLabel.

**These are in very approximate order — many of the steps can be done simultaneously:**

1. Choose your dataset(s). Choose the dataset(s) you plan to make open.
2. Apply an open license. Determine what intellectual property rights exist in the data.
3. Make the data available. In bulk and in a useful format.