FRAUD DETECTION ALGORITHMS WITH AI SOLUTION

SYLLABUS

Dealing with Credit Card Fraud, Machine learning for fraud detection, Fraud Detection and prevention systems, Expert driven Predictive models, Big data analytics in Credit card Fraud detection. Ensemble Learning- Bagging, Boosting algorithms, stacking. IBM Watson Cloud Solutions-Cloud Computing Advantages, Achieving Data Scalability, Cloud delivery models.

- In recent years, we have witnessed an increase in fraudulent activities in the financial sector, and particularly in the area of credit card frauds.
- This is due to the fact that it is rather easy for cybercriminals to set up credit card fraud, and it has, therefore, become important for financial institutions and organizations to be able to promptly identify fraud attempts.
- Furthermore, the activity of fraud detection and prevention in the context of credit card fraud is complicated by the fact that this type of fraud assumes global characteristics; that is, it involves different geographical areas as well as a variety of financial institutions and organizations.
- Therefore, it is essential to be able to share the information sources that are available within different organizations around the world.
- These sources of information are heterogeneous and characterized by explosive growth in data generations, which need to be analyzed in real time.
- This resembles a typical big data analytics scenario, which requires analysis tools and appropriate software and hardware platforms, such as those offered by cloud computing.
- The complexity of the scenario is aggravated by the fact that we are more likely than ever to find money laundering and illegal activities, such as international terrorism financing, to be associated with credit card fraud.
- Illicit activities that are conducted by cybercriminals, therefore, takes on a transnational dimension that involves different sectors of organized crime. All organizations, both in public and private sectors, are called upon to cooperate and counter these illicit activities on the basis of regulatory laws such as anti-money laundering legislation.
- The growing interest of cybercriminals toward credit card fraud is due to distorted economic incentives; the expected payout of credit card fraud is considerably higher than alternative illegal activities, combined with the fact that the risk of being caught by the police is much lower than other forms of traditional crimes.
- Moreover, if individual financial fraud involves amounts of money and values that do not exceed certain thresholds, financial institutions themselves are

discouraged from pursuing illegal activities because investigation activities can prove to be uneconomical

- Financial losses due to credit card fraud are not the only problem that financial institutions must face; there are also reputational damages that are caused by the loss of credibility and reliability.
- Furthermore, credit card fraud can also be a threat to customers; one of the most disturbing aspects of credit card fraud is related to the growing phenomenon of identity theft, which can be easily achieved by creating counterfeit documents or through the appropriation of digital copies of identity documents

DEALING WITH CREDIT CARD FRAUD

Financial institutions have introduced fraud prevention measures over time: in fact, financial institutions have introduced security measures based on two-factor authentication, which integrates traditional authentication procedures by sending an OTP code via SMS to the customer's mobile phone number to prevent abuse in the use of payment instruments. However, the fact remains that such measures are not sufficient and the monetary losses that financial institutions suffer as a result of credit card frauds are still in the order of billions of dollars; therefore, the most effective prevention activities to reduce these losses are procedures based on fraud detection and prevention. The field of analysis associated with credit card fraud detection and prevention is rather complex and will offer us the opportunity to see, in action, different analysis approaches that make use of the techniques of predictive analytics, ML, and big data analytics.

Dealing with credit card fraud is a complex area, increasingly reliant on advanced technologies due to the rise and ease with which cybercriminals perpetrate such activities. Despite measures like two-factor authentication, financial institutions still face significant monetary losses.

- Dealing with fraud involves two distinct activities:
  - **Fraud Detection**: Procedures that identify fraud *after* it has occurred. This involves classifying fraud based on available data.
  - **Fraud Prevention**: Procedures aimed at effectively *preventing* fraud before it happens. This can use rule-based systems or advanced techniques like data mining, ML, and neural networks to find patterns.
- Both detection and prevention aim to **minimise false positives** (legitimate transactions flagged as fraudulent) and avoid denying service to legitimate customers.
- Predictive models for fraud detection follow two main approaches:
  - **Expert-driven (Rule-based)**: Based on rules defined by experts (e.g., `if...then...else`) to represent fraud scenarios. These include

scoring rules to estimate fraud probability and blocking rules to deny transactions. They are easy to implement, understand, and explain, but are subjective, limited in handling complex correlations, based on past experience, and require manual updates.

  - **Data-driven**: Use algorithms from statistics, data mining, and ML to learn hidden patterns from data. ML is key here, identifying models from training data. ML can analyse **multidimensional datasets**, correlate features, dynamically update models, and use large amounts of data in real time. They are generally more robust and scalable. However, they can be "black boxes," making alert justification difficult, and face challenges with data issues like being unbalanced or non-stationary.

MACHINE LEARNING FOR FRAUD DETECTION

The introduction of algorithmic procedures for fraud detection in the credit card sector represents an important test bench in the field of predictive analytics.

The choice and design of appropriate algorithms for credit card fraud detection are characterized by the following:

- Data concerning fraud transactions is not commonly available as financial institutions are reluctant to disseminate such information for fear of reputational damage, as well as confidentiality compliance requirements.
- From a technical point of view, the data on fraud usually represents non- stationary distributions, that is to say, they undergo changes over time; this is also due to the change in customers' spending behaviors.
- Transaction distributions are heavily unbalanced as fraud usually represents a small percentage of overall transactions; therefore, the distributions show a high skewness toward genuine transactions. In fact, we are usually only able to measure fraud that has actually been detected, while it is much more difficult to estimate the number of fraud instances that haven't been detected at all (false negatives). Furthermore, fraud is usually recorded long after it actually occurred.

These intrinsic characteristics of misrepresentations concerning fraud transactions result in challenges in the selection and design of detection and prevention algorithms, such as:

- The use of sampling strategies in data analysis; in the presence of unbalanced distributions the choice of an undersampling/oversampling strategy can be more useful.

- Integration of feedback generated by human operators in identifying fraud alerts. This aspect is particularly important for improving the learning process of algorithms in the presence of non-stationary data, which evolves over time.

All this translates into the development of a fraud detection and prevention system, able to integrate big data analytics, ML algorithms, and human operator's feedback. Therefore, it is clear that the use of cloud computing architectures is the obligatory implementation of choice.

FRAUD DETECTION AND PREVENTION SYSTEMS

There are various possible credit card fraud scenarios, including the following:

- **Theft of credit cards**: This is the most frequent case in practice; criminals steal or spend as much money as possible in a short time span. This activity is noisy and can be identified by means of anomalous or unusual pattern detection that's carried out with respect to the spending habits of the legitimate credit card holder.
- **Credit card abuse**: Unlike the previous case, fraudsters don't need to physically hold the credit card, but it is sufficient that they know the relevant information associated with the card (identification codes, PIN, personal identifier number, card number, device code, and so on). This is represented by one of the most insidious fraud scenarios as it is conducted in stealth mode (it isn't noisy, compared to the previous scenario) and the legitimate owner of the card is often unaware of the ongoing fraud taking place behind his/her back.
- **Identity theft**: In this case, the credit card is issued on the basis of false personal information, or by exploiting the personal information of unsuspecting third parties, who find themselves charged for service costs and withdrawals and payments that have been made in their name.

We should bear in mind that fraud scenarios evolve over time in relation to process and product innovations concerning financial services and technologies that are adopted by financial institutions. Similarly, fraudsters adapt their behavior based on the technical measures that are adopted by credit card issuers to prevent and combat fraud.

- Dealing with credit card fraud involves two distinct activities that comprise the FDPS:
    - **Fraud Detection**: Procedures aimed at correctly and reliably identifying cases of fraud **after** they have occurred. This involves classifying fraud based on available data.

    - **Fraud Prevention**: Procedures aimed at effectively **preventing** fraud **before** it happens. This can be achieved using rule-based alarm systems defined by experts or by leveraging advanced techniques like **data mining, ML, and neural networks** to automatically discover patterns within data distributions.
- Both fraud detection and prevention activities share the goal of **minimising false positives**, which are legitimate transactions incorrectly flagged as fraudulent. This is crucial to avoid denying service to legitimate customers. Managing false positives is challenging due to the poor scalability of manual human checks, highlighting the need for automated procedures to support human operators.
- Implementing effective FDPS procedures faces several challenges related to the nature of fraud data:
    - Data on fraud transactions is often **not commonly available**.
    - Fraud data typically shows **non-stationary distributions**, meaning it changes over time, partly due to evolving customer behaviour.
    - Transaction distributions are **heavily unbalanced**, with fraud representing only a small percentage of overall transactions, resulting in a high skewness towards genuine transactions.
    - It is difficult to accurately estimate the number of **undetected fraud instances (false negatives)**.
    - Fraud is usually recorded considerably after it actually occurs.

EXPERT DRIVEN PREDICTIVE MODELS

The **expert-driven approach**, also known as the **rule-based approach**, implements predictive models based on rules established by experts in a specific domain. Historically, this method is linked to early attempts at automated learning through **expert systems**. An expert system is a computer program using AI technologies to simulate the judgment and behaviour of human experts.

In expert systems, rules were defined to cover all possible cases in a given application domain, with these options being hardcoded and verified by experts. These rules follow logical conditions, typically in the form of `if...then...else` statements, aiming to represent different scenarios and define automatic countermeasures based on data checks.

For credit card fraud detection and prevention systems (FDPS), expert-driven models can be used for fraud prevention procedures, utilising rule-based alarm systems processed by experts. These systems require constant manual fine-tuning. Examples of rules include identifying transactions exceeding a certain amount and daily frequency compared to historical habits, or blocking transactions far geographically from the last one within a short time frame. These can be categorised into **scoring rules** (estimating fraud probability) and **blocking rules** (denying transactions based on stringent conditions).

Advantages associated with rule-based predictive models include:

- **Ease of alerts implementation**.
- **Ease of alerts understanding**.
- **Greater alerts explicability**.

However, expert-driven models have significant disadvantages:

- They express **subjective judgments** and may differ based on the implementing experts.
- They typically handle **only a few significant variables** and their mutual correlations.
- They are **based on past experiences** and **cannot automatically identify new fraud patterns**.
- They require **constant, manual fine-tuning** by experts to keep up with evolving fraud strategies.
- They require manual revisions to incorporate feedback from human operators.

Early expert systems had a fundamental limitation in that they reduced decisions to Boolean values, restricting their ability to adapt to the nuances of real-world use cases. Statistical models, similar to rule-based decisions, also remained rigid as they were established in advance and couldn't adapt to new data.

Data-Driven Predictive Models

- To overcome the limitations of rigid, predefined models like those in expert systems and static statistical models, it became necessary to adopt an iterative approach. **Data-driven predictive models** address this by exploiting **automated learning algorithms**.
- These models attempt to adapt their predictions based on learning directly from data, constantly updating detection and prevention procedures based on dynamically identified behaviour patterns.
- The algorithms used are derived from fields like statistics, data mining, and Machine Learning (ML), with the objective of learning hidden or latent patterns within the data. ML plays a privileged role here, enabling the identification of predictive models based on data training.
- This approach allows algorithms to generalize descriptive models from available data, autonomously generating features and adapting to the continuous evolution of the training process.
- Data mining, the discovery of adequate representative models starting with the data, reflects this difference in approach from predefined static models.

- When the nature of data is clear and conforms to known models, using pre-defined models might suffice, but the ability to manage cases not covered in training data leads to the adoption of AI, which includes ML.

Advantages of using ML in data-driven predictive models include:

- The ability to analyze **multidimensional datasets** (those with many features).
- The ability to **correlate features** with each other.
- The ability to **dynamically update models**, adapting them to changes in strategies adopted by fraudsters.
- The use of a data-driven approach that leverages **large amounts of data (big data) in real time**.

Data-driven predictive models are generally **more robust and scalable** than rule-based models. They have the undoubted advantage of being able to **integrate feedback generated by human operators** automatically, improving prediction accuracy and reducing false negatives.

However, data-driven models often behave like **black boxes**, meaning the alerts they generate can be difficult to interpret and justify. Also, the nature of data, such as it being **unbalanced, non-stationary, and skewed**, can pose difficulties in implementing algorithms correctly.

Predictive Analytics for Credit Card Fraud Detection

Credit card fraud represents a crucial application area for AI solutions in cybersecurity. The significant monetary losses from fraud, despite measures like two-factor authentication, highlight the need for effective fraud detection and prevention procedures. The field of analysis for credit card fraud detection and prevention is complex, making use of techniques from **predictive analytics, ML, and big data analytics**.

A comprehensive **Fraud Detection and Prevention System (FDPS)** requires the integration of big data analytics, ML algorithms, and feedback from human operators. Dealing with credit card fraud involves two main activities:

1. **Fraud Detection**: Identifying fraud cases **after** they have occurred, primarily through classifying fraud based on available data.
2. **Fraud Prevention**: Effectively **preventing** fraud **before** it happens. This can be achieved using rule-based alarm systems defined by experts (requiring fine-tuning) or by leveraging advanced techniques like data mining, ML, and neural networks to automatically discover data patterns.

Both detection and prevention activities share the critical need to **minimise false positives** (legitimate transactions flagged as fraudulent) to avoid denying service to customers. Managing false positives manually is difficult due to poor scalability, necessitating automated procedures to support human operators.

Developing predictive analytics models for credit card fraud detection involves tackling several challenges due to the nature of fraud data:

- Data on fraud transactions is **not commonly available**.
- Fraud data often shows **non-stationary distributions**, changing over time partly due to evolving customer behaviour and fraudsters adapting strategies.
- Transaction distributions are **heavily unbalanced**, with fraud being a small percentage of overall transactions, resulting in high skewness towards genuine transactions.
- It is difficult to accurately estimate the number of **undetected fraud instances (false negatives)**.
- Fraud is usually recorded considerably after it actually occurs.

These challenges necessitate specific strategies in the design and selection of algorithms, including:

- Using **sampling strategies** like undersampling or oversampling (such as SMOTE) to manage unbalanced data distributions.
- **Integrating feedback generated by human operators** to improve algorithm learning, particularly important for non-stationary data and reducing false negatives. Data-driven models can integrate this feedback automatically, unlike rule-based models which require manual revisions.

Predictive analytics for fraud detection applies data mining and ML techniques to large amounts of data (big data analytics) to identify trends. It aims to predict future events by extrapolating hidden patterns, contrasting with descriptive analytics which only classifies past data.

BIG DATA ANALYTICS IN CREDIT CARD FRAUD DETECTION

- Credit card fraud detection and prevention is a complex area that incorporates techniques from predictive analytics, machine learning (ML), and **big data analytics**.
- It is a significant domain for applying AI solutions in cybersecurity, requiring predictive analytics models that leverage **big data analytics**, often facilitated by cloud computing platforms.
- The nature of credit card fraud, involving widespread activities across geographies and institutions, combined with the heterogeneous and rapidly

increasing volume of data generated, aligns closely with a typical **big data analytics** scenario.

- Handling this volume and diversity of data in real-time demands specialised analysis tools and appropriate software/hardware platforms, such as those provided by cloud computing.
- Predictive analytics for fraud detection is fundamentally data-driven, relying on data mining and ML techniques applied to vast datasets, characteristic of **big data analytics**.
- Traditional data management systems like relational databases and data warehouses used for standard reporting (descriptive analytics) are often inadequate for the scale and data-driven nature of predictive analytics in this field.
- Adopting data architectures that offer processing scalability through functional programming paradigms like MapReduce and NoSQL primitives is necessary for managing this data.
- **Big data analytics** techniques can be integrated with ML and data mining algorithms to automate the process of detecting fraud.
- Embracing the **big data analytics** paradigm allows organisations to maximise the value from their information assets, even when they originate from diverse and heterogeneous sources.
- This capability enables the implementation of advanced forms of **contextual awareness**, allowing fraud detection procedures to adapt to changing contexts in real-time.
- For instance, integrating internal data with publicly available data from sources like websites and social media, facilitated by cloud platforms, helps reconstruct the context of financial transactions. Social media data might reveal a cardholder's location, which can be cross-referenced with the transaction location.
- The integration of various data sources, made possible by **big data analytics**, also supports **feature augmentation**. This involves creating new variables (features) from existing ones to better describe the behaviour of legitimate cardholders and differentiate it from fraudulent activity. Examples include metrics like average expenditure, daily purchase frequency, and common purchase locations. This keeps customer profiles current and helps detect anomalies promptly.
- The growth of **big data analytics** has driven the adoption of distributed storage systems, known as NoSQL databases, to prevent bottlenecks in data management and storage. Cloud computing extensively utilises these systems for analysing large datasets, including streaming data.
- A comprehensive fraud detection and prevention system (FDPS) capable of effectively addressing credit card fraud requires the integration of **big data analytics**, ML algorithms, and feedback from human operators. Cloud computing architectures are considered the necessary implementation choice for such systems.

- While data-driven technologies like AI and **big data analytics** are powerful, their use of personal data in areas like fraud detection raises significant issues regarding data security and confidentiality. Aggregating personal data can lead to the creation of highly probable individual profiles. Using automated analysis of these profiles for business decisions carries the risk of incorrect profiling, potentially resulting in negative discrimination and legal issues. Maintaining the integrity and confidentiality of the data used by algorithms is crucial for their security and reliability.

ENSEMBLE LEARNING

- The **purpose of ensemble learning** is to combine different classification algorithms to create a classifier that provides **better predictions** than any single individual classifier.
- In the presence of non-stationary data (data whose statistical properties change over time), using an ensemble of classifiers can be useful to **improve overall prediction accuracy**.
- By using combinatorics analysis and binomial probability distribution, it is possible to show that combining multiple binary classifiers can **improve the probability of obtaining correct predictions** and **reduce the probability of errors**, even if individual classifiers have an error rate. For example, if 11 binary classifiers with a 25% error rate are used together, the error rate can be reduced to 3.4%.
- One method for combining classifiers is **majority voting**, also known as the majority voting principle. This involves selecting the prediction among the individual classifiers that has the highest frequency, which formally translates to calculating the statistical measure of position known as the mode.
- Individual classifiers within an ensemble can be chosen from **different types of algorithms**, such as decision trees, random forests, and support vector machines (SVMs).

Bagging

- Bagging stands for "**bootstrap aggregating**".
- The term "bootstrap" refers to the operation of **sampling with replacement** applied to a dataset.
- The bagging method associates an individual estimator (classifier) with each bootstrap sample.
- The ensemble estimator is created by applying a method like **majority voting** to the individual classifiers trained on these bootstrap samples.
- Bagging is particularly useful for **reducing the variance of individual estimators** by selecting different training sets.

- It is especially beneficial when sampling with replacement helps to **rebalance the original dataset**, thereby reducing total variance.
- The scikit-learn library provides the `BaggingClassifier` class for implementing this method. Parameters like `n_estimators` (number of base estimators) and `max_samples` (maximum number of samples per base estimator) can be set, and the `bootstrap` parameter can be set to `True` to activate the bootstrap mechanism.

Boosting

- The boosting method uses **weighted samples** extracted from the data.
- The weights of these samples are **readjusted iteratively** based on the classification errors reported by the individual classifiers.
- Greater importance (weight) is given to observations that were more difficult to classify.
- The primary goal of boosting is to create an ensemble estimator that **reduces the bias of the individual classifiers**.
- One of the best-known boosting algorithms is **Adaptive Boosting (AdaBoost)**. AdaBoost trains a first classifier, increases the weight of incorrectly classified samples, trains a second classifier on the updated dataset, and continues this sequential process until a predetermined number of estimators is reached or an optimal predictor is found.
- A main disadvantage of AdaBoost is its **sequential learning strategy**, which prevents parallel execution.
- **Gradient Boosting** is another boosting method available in scikit-learn via the `GradientBoostingClassifier` class. Default estimators in Gradient Boosting are often decision trees. The `learning_rate` parameter in Gradient Boosting determines the contribution of each estimator to the ensemble classifier; a low value requires more estimators.
- Scikit-learn provides implementations like `AdaBoostClassifier` and `GradientBoostingClassifier`.

Stacking

- The stacking method constructs the ensemble estimator by **superimposing layers**.
- The **first layer consists of individual estimators** (base classifiers).
- The **predictions** from the first layer's estimators are then forwarded to a **second layer**.
- In the second layer, **another estimator** (often called a meta-estimator) is tasked with classifying the predictions received from the first layer.

- Unlike bagging and boosting, stacking can utilise **different types of basic estimators** in the first layer, and the final estimator in the second layer can also be of a different type than the basic estimators.

eXtreme Gradient Boosting (XGBoost)

An algorithm that's similar to gradient boosting is the XGBoost algorithm. It represents an extension of gradient boosting that proves to be more suitable in managing large amounts of data since it is more scalable. XGBoost also uses the gradient descent method to minimize the residual error of the estimators, and is particularly suitable for parallel computing (a feature that makes it more suitable for cloud computing).

Sampling methods for unbalanced datasets

The two most adopted sampling modes are undersampling and oversampling. Through undersampling, some random samples are removed from the most numerous class (in our case, the class of legitimate transactions); with oversampling, synthetic samples are added to the class with the lowest occurrences.

Oversampling with SMOTE

Among the oversampling methods, we have the Synthetic Minority Over-sampling Technique (SMOTE); this allows for the generation of synthetic samples by interpolating the values that are present within the class subjected to oversampling. In practice, synthetic samples are generated based on the clusters that are identified around the observations present in the class, therefore calculating the k-Nearest Neighbors (k- NNs). Based on the number of synthetic samples that are needed to rebalance the class, a number of k-NN clusters are randomly chosen, around which synthetic examples are generated by interpolating the values that fall within the selected clusters.

IBM WATSON CLOUD SOLUTIONS

- The **IBM Watson Cloud solution** is presented as one of the interesting cloud-based solutions available for implementing **credit card fraud detection**. This concrete example of fraud detection highlights the use of the **IBM Cloud platform**.
- The IBM Watson Cloud solution also introduces the innovative concept of **cognitive computing**. Through cognitive computing, it is possible to **emulate**

the typically human ability of pattern recognition, which allows for obtaining **adequate contextual awareness for decision-making**.

- IBM Watson can be successfully used in various real-world scenarios, including:
  - Augmented reality
  - Crime prevention
  - Customer support
  - Facial recognition
  - **Fraud prevention**
  - Healthcare and medical diagnosis
  - IoT
  - Language translation and natural language processing (NLP)
  - **Malware detection**
- Credit card fraud detection is described as requiring predictive analytics models that exploit **big data analytics through the use of cloud computing platforms**. The scenario of managing fraud detection often involves amounts of data that cannot be effectively analysed with traditional ETL procedures on relational databases, necessitating scalable AI solutions and cloud architectures to manage big data and predictive analytics.
- The use of cloud computing architectures is considered the **obligatory implementation of choice** for fraud detection and prevention systems that integrate big data analytics, ML algorithms, and human operator feedback.
- **Cloud computing** has rapidly gained prominence due to higher bandwidth networks and low-cost computing/storage, enabled by virtualisation solutions. Its central element is **scalability**.
- Advantages of adopting cloud computing solutions include:
  - Optimising IT investments and improving profit margins.
  - Benefiting from an **on-demand model** for resources, reducing fixed costs and converting them to variable costs.
  - Efficiently storing and managing **large amounts of data** (big data analytics).
  - Guaranteeing **high performance, high availability, and low latency**.
  - Storing and replicating data on geographically distributed servers.
  - Enabling scalability through data partitioning and managing increasing workloads by adding resources linearly.
- Cloud computing makes extensive use of **distributed storage systems** (NoSQL databases) to prevent bottlenecks in data management, allowing data analysis even in streaming mode. These systems store data in key-value pairs and enable parallel processing using functional programming paradigms like MapReduce, leveraging the distributed computing capabilities of the Cloud. NoSQL databases also offer flexible data management without needing to reorganise the structure as analysis changes.

- The IBM Cloud platform offers a delivery model including **Infrastructure as a Service (IaaS)** and **Platform as a Service (PaaS)**. It also provides a series of cloud services that can be integrated into applications, such as:
    - **Visual recognition**: Locating information like objects, faces, and text in images/videos, with options for pre-trained or custom models.
    - **Natural language understanding**: Extracting sentiment and identifying information (people, places, organisations, concepts, categories) from text, useful for analysing sources like social media to contextualise events like credit card transactions. This service can be adapted using Watson Knowledge Studio.
- The IBM Cloud platform also offers advanced tools for application development:
    - **Watson Studio**: Manages projects, facilitates team collaboration, adding data sources, creating Jupyter Notebooks, training models, and provides data analysis features like data cleansing.
    - **Knowledge Studio**: Allows the development of customized models based on specific company needs, usable by Watson services.
    - **Knowledge Catalog**: Manages and shares company data, supports data cleaning/wrangling, and profiles data access permissions through security policies.
- A major advantage of the IBM Cloud platform is the ability to implement advanced solutions leveraging cognitive computing.

CLOUD COMPUTING ADVANTAGES

- With the widespread availability of **higher bandwidth networks**, combined with **low-cost computing and storage**, the architectural model of cloud computing has rapidly become prevalent thanks to virtualisation solutions.
- The **central element** characterising cloud computing is its **scalability**, which has contributed significantly to its commercial success.
- Organisations that have adopted cloud computing solutions have successfully **optimised investments in IT**, leading to **improved profit margins**.
- Instead of needing to size their technological infrastructure for the worst-case scenario (accounting for temporary workload peaks), organisations using cloud solutions benefit from an **on-demand model**, converting fixed costs into **variable costs**.
- This improvement in IT investment quality allows organisations to focus on the **management and analysis of their data**.
- Cloud computing enables the **efficient storage and management of large amounts of data** (big data analytics).
- It guarantees **high performance**, **high availability**, and **low latency**.
- To ensure access guarantees and performance, data is **stored and replicated on servers distributed across various geographical areas**.

- By partitioning data, benefits related to the **scalability of the architecture** are achieved.
- Scalability specifically relates to the ability to **manage increasing workloads by adding resources**, with costs increasing linearly proportional to the added resources.
- Cloud computing extensively uses **distributed storage systems** (such as NoSQL databases storing data in key-value pairs) to **prevent bottlenecks** in data management and storage, enabling the analysis of large amounts of data, even in streaming mode.
- Distributed storage systems allow for **parallel data processing** using functional programming paradigms like MapReduce, fully leveraging the distributed computing capabilities offered by the Cloud.
- The use of NoSQL databases also provides **flexible data management**, removing the need to reorganise the entire archive structure when analysis changes. This is beneficial when verifying predictive model accuracy in real-time for decision-making, particularly in cybersecurity.
- The scalability and on-demand resource management capability allow providers to offer different cloud delivery models, including Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS).
- The IBM Cloud platform offers a delivery model including **IaaS and PaaS**, along with cloud services like Visual Recognition and Natural Language Understanding that can be integrated into applications.
- The IBM Cloud platform offers advanced development tools like **Watson Studio** (for project management, collaboration, data sources, Jupyter Notebooks, model training, data analysis/cleansing), **Knowledge Studio** (for developing custom models for Watson services), and **Knowledge Catalog** (for managing/sharing company data, cleansing/wrangling, profiling data access permissions).
- A major advantage of the IBM Cloud platform is the ability to implement advanced solutions leveraging **cognitive computing**.

ACHIEVING DATA SCALABILITY

Achieving **data scalability** is a fundamental aspect of managing large datasets, particularly in modern scenarios like **big data analytics** and applications such as fraud detection and cybersecurity. Traditional architectures based on relational databases and data warehouses do not scale well when faced with explosive data growth.

Here are the key ways achieving data scalability:

- **Adoption of Cloud Computing Architectures** The use of **cloud computing architectures** is considered the **obligatory implementation of choice** for systems that integrate big data analytics, ML algorithms, and human operator feedback. Cloud platforms, such as **IBM Cloud** solutions, are specifically presented as allowing companies to take advantage of cloud architectures to manage big data and leverage predictive analytics methodology.
- **Leveraging the Scalability of Cloud Computing** The central element characterising cloud computing is its inherent **scalability**. Scalability relates to the ability to manage increasing workloads by adding resources to the architecture, with costs increasing linearly proportional to the added resources. Cloud solutions provide an **on-demand model** for resources, which allows organisations to benefit by converting fixed costs into variable costs, rather than having to size their infrastructure for worst-case scenarios.
- **Efficient Storage and Management of Large Data** Cloud computing allows for the efficient storage and management of **large amounts of data** (big data analytics). To guarantee access and performance, data is stored and replicated on servers distributed across various geographical areas.
- **Use of Distributed Storage Systems (NoSQL Databases)** With the spread of big data analytics, it became necessary to move to **distributed storage systems** to prevent bottlenecks in data management and storage. Cloud computing makes extensive use of these systems, which are non-relational databases defined as **NoSQL databases**, storing data in key-value pairs.
- **Functional Programming Paradigms** Distributed storage systems allow for the management of data in a distributed mode on multiple servers by following **functional programming paradigms** such as **MapReduce**. This enables the execution of data processing in parallel, fully leveraging the distributed computing capabilities offered by the Cloud. These architectures allow the achievement of processing scalability.
- **Flexible Data Management with NoSQL** The use of NoSQL databases also allows data to be managed in a flexible manner, without the need to reorganise its overall structure as the analysis changes. This is particularly important for verifying predictive model accuracy in real time in cybersecurity.
- **Data Partitioning** By **partitioning the data**, it is possible to obtain the advantages connected to the scalability of the architecture.
- **Integrating Heterogeneous Data Sources** Embracing the paradigm of **big data analytics** helps organisations make the most of their information assets from different (often heterogeneous) data sources. Integrating different data sources allows for feature augmentation of datasets, introducing new variables to describe behaviour. This integrated approach helps in implementing advanced forms of contextual awareness and adapting detection procedures in real time. Fraud detection requires the integration of heterogeneous data sources.
- **Adopting Scalable AI/ML Algorithms** Scalable AI solutions are required to manage the dimensions of data that cannot be effectively analysed with

traditional methods. The scalability of algorithms is an essential element to improve performance, especially when deploying solutions into the cloud. While not all algorithms are designed to guarantee scalability, data-driven predictive models are usually more robust and scalable than rule-based models. Algorithms like XGBoost are more suitable for managing large amounts of data and parallel computing, making them more suitable for cloud computing.

- **Reducing Dataset Dimensionality** Reducing the dataset dimensionality, meaning the number and type of features used, can dramatically improve the performance of algorithms. Techniques like Principal Component Analysis (PCA) can be used for dimensionality reduction of large datasets.

CLOUD DELIVERY MODELS

The **scalability of the architecture**, coupled with the capability to manage resources in an **on-demand mode**, allows cloud providers to offer various **cloud delivery models**.

These models represent different service offerings with varying levels of abstraction and management handled by the provider. There are three primary models:

- **Infrastructure as a Service (IaaS)**: In this model, the provider deploys the foundational **IT infrastructure**. This infrastructure includes core components such as **storage capabilities** and **networking equipment**. With IaaS, users typically manage the operating systems, applications, and data, while the provider handles the underlying hardware.
- **Platform as a Service (PaaS)**: With PaaS, the provider offers a higher level of abstraction by deploying necessary elements for application development and deployment, including **middleware**, a **database**, and other required components. PaaS provides a ready-to-use platform, allowing users to focus on writing and deploying their applications without managing the underlying infrastructure or operating systems.
- **Software as a Service (SaaS)**: This model is the most abstracted, with the provider deploying and managing **complete applications**. Users access the software over the internet, often via a web browser, and do not need to install, manage, or update the application or its underlying infrastructure.

The **IBM Cloud platform** is highlighted as offering a delivery model that encompasses both **IaaS and PaaS**. In addition to these core infrastructure and platform layers, IBM Cloud also provides a range of **cloud services** that can be integrated into applications developed by organisations. These integrated services support various functionalities, including:

- **Visual Recognition**: This service allows applications to **locate information** within images and videos, such as identifying objects, faces, and text. The platform offers pre-trained models and the ability to train models using corporate datasets.
- **Natural Language Understanding (NLU)**: This service enables the extraction of information from text, including analysing **sentiment**. It is particularly useful for extracting information from sources like social media to gain insights, such as whether a credit card holder is in a foreign country during a transaction. The service can identify entities like people, places, organisations, concepts, and categories, and can be adapted for specific application domains using Watson Knowledge Studio.

The IBM Cloud platform also offers a series of advanced tools for application development:

- **Watson Studio**: This allows the management of projects and offers tools for collaboration between team members. With Watson Studio, it is possible to add data sources, create Jupyter Notebooks, train models, and use many other features that facilitate data analysis, such as data cleansing functions. We will have the opportunity to deepen our knowledge of Watson Studio soon.
- **Knowledge Studio**: This allows the development of customized models on the specific needs of the company; once developed, the models can be used by Watson services, in addition to, or in place of, the predefined models.
- **Knowledge Catalog**: This allows the management and sharing of company data; the tool also makes it possible to perform data cleaning and wrangling operations, thereby profiling data access permissions through security policies.

A significant advantage offered by the IBM Cloud platform is the capability to implement advanced solutions by leveraging **cognitive computing**. These integrated services and tools support this capability, allowing for sophisticated data analysis and application development.