UNIVERSITÀ DEGLI STUDI DI CATANIA
Department of Economics and Business
MSc in Data Science for Management

Isola Catania: Transforming Data into Vision

Manuel Scionti

Supervisor: Prof. Salvatore Ingrassia
Co-supervisors: Prof. Antonio Punzo
Dr. Luca Naso

Academic year 2022 - 2023

Abstract

In the ever-evolving landscape of modern businesses, data-driven decision-making has become paramount for organizations to thrive and adapt. This Master's thesis delves into a real-world case study, centered on assisting an innovation hub and co-working space in initiating its data-driven journey. The study is divided into two pivotal tasks: the construction of a comprehensive DataLake utilizing Amazon Web Services (AWS) to consolidate diverse data sources and cluster analysis to facilitate user segmentation and profiling of their user base.

I learned to walk: ever since, I let myself run.
— F.W. Nietzsche

Acknowledgments

La fine di un percorso, di un viaggio. Un capitolo della mia vita che si conclude e chissà quanti altri ancora da scrivere.

Ringrazio calorosamente il mio relatore, il Professore Ingrassia, e i miei correlatori, il Professore Punzo e il Dr. Luca Naso. Vi ringrazio per aver avuto fiducia in me e avermi guidato lungo tutto questo percorso. La vostra competenza e il vostro supporto sono stati fondamentali per la riuscita di questo progetto.

Ti ringrazio Luca per avermi accolto in Koexai e avermi fatto sentire fin da subito parte della grande e bellissima squadra che hai creato. Ciò che mi avete trasmesso tu e Giovanni non lo dimenticherò mai e faranno per sempre parte del professionista che diventerò.

Ringrazio Antonio Perdichizzi e tutte le ragazze e i ragazzi di Isola. Grazie per esservi messi in gioco e aver creduto in me. Vi ringrazio per il vostro spirito, per il vostro amore verso la nostra bella ma amara terra. Grazie per voler lasciare il mondo un posto migliore di come l'avete trovato.

Un ringraziamento speciale va alla mia famiglia, a mia madre, a mio padre, al mio amato fratellino e a Mario. Senza di voi non sarei la persona che sono adesso. Grazie per avermi sempre supportato e sopportato. E grazie soprattutto perchè so che continuerete a farlo sempre, qualsiasi cosa accada. Grazie ai miei amici di sempre, grazie Riccardo, Alessia, Simone, Stefano, Angelo, Giuseppe. Presenza costante e fondamentale in tutti questi anni di follie. Un grazie alle nuove e alle vecchie amicizie, alla famiglia di ESN e ai colleghi con cui ho condiviso questi due anni indimenticabili.

Grazie M., non sai quanto sono orgoglioso di te e grato per la persona che mi hai reso. Ogni passo ci avvicina alla cima, dove spero ci ritroveremo.

Zia, un bacione da quaggiù.

Contents

List of Figures

List of Tables

List of Equations

Introduction

In today's highly competitive and fast-changing business world, making decisions based on data has become essential for sustainable success. The digital revolution has affected all sectors and led to the creation of enormous amounts of information, including details on consumer actions, operational measurements, business trends, and beyond. By studying and interpreting this information, businesses can gain a better understanding of complex situations. Relying solely on old-fashioned decision-making methods based on past experiences, intuition or anecdotal evidence can make businesses shortsighted and less flexible. However, a data-focused approach offers a more complete, unbiased and future-facing view. It helps companies to spot new trends, acknowledge hidden market demands, manage resources efficiently, and anticipate and prevent risks. In active places such as innovation hubs and co-working environments, where a wide variety of professionals, start-ups and companies converge, this is particularly important. A mix of different approaches and plans is essential in these places. Managers and decision-makers can obtain valuable insights by using the data from these areas, which can enable them to personalize services, create focused programmes, and foster a favourable environment for innovation and cooperation. Additionally, in a constantly evolving era of customer and user anticipations, data-generated insights function as a guide for enterprises to conform their strategies according to genuine needs. In the current business world, data is key to making informed decisions, being strategically agile, and achieving sustained growth. The aim of this thesis is to document the progression of Isola Catania towards a more data-oriented management and investigate its benefits. Koexai s.r.l. has been a valuable collaborator throughout this journey, providing their expertise and support from the outset. Their knowledge has proved instrumental in helping Isola achieve its ambitions. Therefore, this research serves as a factual illustration and effective demonstration detailing how a nascent enterprise or startup can restructure its administration to adopt data-focused methods, enhance its assets and acquire key insights into its customer base.

The dissertation follows a systematic approach, beginning with a basic overview that sets the research in the context of innovation hubs and co-working spaces. Chapter 1 sets the research question and its proposed solution and presents the main partners of this study. After this introduction, Chapter 2 moves into a comprehensive literature review. It aims to give a comprehensive view of the topic. Chapter 3 focus on practical methods, explaining the complex process of collecting and preparing data. These include detailed investigations of Isola's data sources, as well as the standards applied for their selection. After laying the groundwork for data collection, Chapter 4 moves on to explain the process of constructing a Data Lake via AWS. This section combines theoretical understanding with hands-on experience, displaying AWS's architecture, motives for service choices, and thoughts on security, compliance, and scalability. The key analytical aspect of the dissertation is the segment devoted to cluster analysis for classifying users in Chapter 5. We examine a specific data source, rigorously process data, develop features, and apply clustering algorithms. We interpret the results in the context of user behavior, preferences, and business implications. After presenting the primary findings, Chapter 6 evaluates the outcomes of constructing Data Lake and performing cluster analysis in a results and discussion section. This chapter encourages thinking critically, aligning findings with the main goals of the innovation hub, and discussing the issues and ideas obtained from the research process. The final sections of this thesis summarize important discoveries, contributions to academic and practical areas related to data-driven decision-making, and possibilities for future research and investigation.

Background and Partners Involved

1.1 Koexai S.r.l.

Figure 1.1: Koexai logo

Koexai S.r.l. is an innovative startup specialized in the field of Data Science, spanning a wide spectrum of sectors including Cloud Computing, Big Data, Artificial Intelligence, Machine Learning, and Data Visualization. The company stands out for its versatile operations, ranging from the development of cloud platforms to scientific projects such as the creation of customized Machine Learning algorithms. In addition, the company offers strategic consulting services and tailored training, adapted to the specific needs of the clients, considering their progress in the transformation process towards a data-oriented business model, known as *data-driven*.

To successfully tackle the challenges of the highly technological and innovative sectors in which it operates, Koexai has assembled a team of highly qualified professionals, graduates, holders of research doctorates, and professional certifications. This team not only ensures a high standard of excellence in every project but also promotes an approach based on continuous learning, crucial in a sector where novelties emerge at least annually. Founded in Catania in 2022, Koexai quickly gained a solid reputation for the implementation of cutting-edge projects both from a technological and scientific point of view, with resonance at a national level. Its dedication to innovation constantly drives it to explore new horizons and develop creative solutions to tackle complex challenges.

1.2 Isola Catania Impresa Sociale S.r.l.

Figure 1.2: Isola logo

Isola Catania Impresa Sociale S.r.l.[1] is an impactful and innovative community hub located in the historical and UNESCO-protected Palazzo Biscari in the heart of Catania. It functions as a coworking space and houses offices, playing a pivotal role in enhancing the quality of life in Sicily. Isola addresses critical issues like educational poverty, unemployment, and low cultural and political participation by fostering projects in entrepreneurship, socio-environmental sustainability, and culture. Its efforts aim to create opportunities for youth, forming a virtuous community of businesses, organizations, and extraordinary locales to support the south of Italy in generating value and talent. Isola's operational facets can be segmented as:

- **Innovation**: It serves as a hub that promotes and supports the birth and growth of entrepreneurial projects, startups, and innovative SMEs.
- **Education**: Isola provides educational opportunities designed to develop and amplify local talents.
- **Coworking**: It operates as an experimentation laboratory for the future of work, providing a home for southworkers aiming to create employment in the southern regions.
- **Culture**: Isola engages in the promotion of arts and culture through its events, workshops, concerts, and exhibitions.

[1] From now on we shall refer to it as Isola

The hub originated from the redevelopment of a disused space within Palazzo Biscari, a unique cultural and historical heritage site. It is equipped with state-of-the-art technologies and offers a range of attractive services nationally and internationally, including business localization, coworking spaces, facilities for digital nomads, and event hosting. It has become a home for a robust local community and acts as a cultural facilitator and attractor in the areas of innovation, entrepreneurship, education, sustainability, and work. Palazzo Biscari, where Isola is located, is an excellent historical building near Piazza Duomo. Reconstructed and expanded after the 1963 earthquake, it is a symbol of late-baroque architecture and represents the identity of Catania. Isola occupies about 1000 square meters within the palace, where contemporary and historical elements coexist, creating a versatile and international atmosphere.

Isola regularly hosts students from around the world, playing a key role in connecting them with the entrepreneurial ecosystem of Catania. For example, it has hosted the "Field Trip" of the ESCP business school, allowing students to explore the local entrepreneurial landscape. The project is dedicated to generating positive social and territorial impact, with a commitment to minimizing environmental effects and supporting initiatives for urban ecosystem protection. Isola aims to obtain environmental certifications to validate the sustainability of its festivals and events. It has set ambitious goals, including supporting the establishment of over 25 new businesses, attracting more than 100 companies to Sicily, creating over 500 new jobs from educational programs, and building a community of at least 1000 Sicilians living abroad by 2030.

1.3 Problem Statement

1.3.1 Research question

Isola's data infrastructure is currently fragmented and siloed, making it difficult to use data effectively. This lack of cohesion makes essential tasks, such as calculating impact indicators and conducting advanced business analyses, including analyses of clients and events, more complex. The existing data collection tools are static and offer limited opportunities for comprehensive data analysis. The isolated nature of data sources at Isola has so far prevented their extensive evaluation. Consequently, this ineffective utilization of internal data leads to a limited understanding of Isola's strengths and weaknesses, as well as an unclear view of their target audience. This situation highlights the need for a more integrated and dynamic approach to data management and analysis.

1.3.2 Proposed solutions

The purpose of this thesis is twofold, aimed at enhancing Isola's data management and user understanding through innovative and practical data science applications. The first objective is the creation of a comprehensive Data Lake on Amazon Web Services (AWS). This initiative is designed to address the challenges posed by the current fragmented data infrastructure at Isola. By consolidating various data sources into a centralized Data Lake, the organization is positioned to manage its data more effectively and efficiently. This unified data repository not only facilitates smoother data management but also unlocks the potential for deriving deeper insights into core activities. These activities encompass the multifaceted dynamics of their coworking space, the meticulous organization of events and workshops, and the nurturing environment provided for startups and third-sector organizations. The implementation of this Data Lake promises a transformative impact on how Isola leverages its data for strategic decision-making and operational excellence.

The second purpose of the thesis is to conduct a detailed cluster analysis of the coworking space user base. This analysis seeks to divide and characterise the varied user community whilst revealing data patterns and preferences. By implementing advanced clustering methods, the investigation aims to shed light on the distinctive characteristics, behaviours and needs of different user groups. This comprehensive knowledge is crucial for optimising Isola's services and offerings, improving user experience, and creating a more involved and content community. These objectives work together in synergy to drive Isola towards a data-driven future, characterised by informed decisions, optimised operations and a deeper connection with its user community.

Literature Review

2.1 The importance of data-driven decision-making in business

In the fast-changing business world of today, data plays a crucial role in shaping the strategies and results of organisations. With the constant flow of information that companies have to handle, the capability of using this data to drive decision-making has become an essential distinguishing factor. The move towards *Data-Driven Decision-Making* (D^3M) is backed by plenty of academic research, which emphasises its significance and offers advice on how to put it into practice. The idea of using data to inform decisions is not a novel one. Companies have always employed information to help shape their plans. But the digital era has brought with it an unprecedented volume, range and speed of data. This surge of information, combined with progress in analysis and computing, has altered how organisations perceive and utilise data. Hartmann et al. (2016) [8] provide a comprehensive examination of this transformation. They emphasise that D3M is not just about the accumulation of large data sets. Instead, it involves the complex integration of data into business processes to improve decision making. This integration, they argue, can give companies a competitive edge, but it requires a harmonious blend of technological, organisational and individual skills.

The term *data-driven organisation* (DDO) has gained considerable traction in recent years. But what does it mean to be truly data-driven? Davenport and Bean (2018) [3] explore this question in depth. They note that a true DDO is one that instils a culture that values and champions the use of data in decision making. This involves more than just investing in the latest analytics tools. It requires a paradigm shift in which employees at all levels understand, value and use data in their day-to-day work. Such an organisational culture, they argue, can lead to improved business outcomes, including increased profitability and superior customer satisfaction. Fischer et al. (2022) [6] further this discourse with a comprehensive review of the evolution and nuances of data-driven organisations. They note that while there's a growing interest in the DDO concept, interpretations of what it entails vary widely. Their study synthesises these different perspectives and offers a conceptual framework for understanding DDOs. This framework identifies five core elements that are integral to a DDO:

1. **Data sourcing & sensemaking**: This refers to how organisations gather data and make sense of it. It involves filtering out noise, identifying patterns and discerning actionable insights.

2. **Data capabilities**: This includes the technological and analytical capabilities that enable organisations to effectively process and analyse data.

3. **Data-Driven Culture**: This refers to the organisational ethos that promotes the use of data in decision making. It's about fostering a mindset where data is seen as a valuable asset.

4. **Data-Driven Decision-Making (D^{3}M)**: This is the actual process of leveraging data to make informed decisions. It involves analyzing data, drawing inferences, and making choices that align with organizational objectives.

5. **Data-Driven Value Creation**: This is about translating data-driven decisions into tangible value – be it in terms of revenue growth, cost savings, or other business metrics.

Fischer et al. [6] argue that an archetypal DDO seamlessly integrates an outside-in perspective (gathering external data and insights) with an inside-out view (leveraging internal data and influencing the external environment). Transitioning to a data-driven paradigm is not devoid of challenges. Hartmann et al. (2016) [8] sound a note of caution. While data can offer invaluable insights, it's imperative to ensure its accuracy and relevance. Erroneous or outdated data can lead to misguided decisions with detrimental consequences. Furthermore, the sheer volume of data can be overwhelming. Filtering out pertinent information from the noise becomes crucial. They also underscore the importance of skilled personnel adept at interpreting and analyzing data. Without the right talent, even the most sophisticated analytics tools can prove ineffectual.

The impact of data-driven decisions on businesses is profound. McAfee and Brynjolfsson (2012) [16] provide empirical evidence to this effect. They note that companies adept at leveraging data witness a marked increase in productivity and profitability. Such companies are also more innovative, leading to the development of novel products and business models. Moreover, being data-driven equips organizations to be more agile and responsive to market shifts and evolving customer needs, bestowing a competitive edge. In other words, data-driven decision-making is revolutionizing the business world. Organizations that adeptly integrate data into their decision-making processes stand to reap rich dividends in terms of operational efficiency, profitability, and market leadership. However, realizing these benefits demands more than just investing in technology. It requires a holistic approach that melds technology, culture, and talent. As businesses embark on this data-driven journey, they must be cognizant of the challenges and navigate them judiciously to harness the true power of data.

2.2 Relevance of data lakes in modern data architectures

In the contemporary era of digital transformation, businesses are increasingly leaning towards data-driven decision-making to maintain a competitive edge. The rise of Big Data has necessitated the evolution of data management strategies, with Data Lakes emerging as a pivotal concept in this landscape. This section delves deep into the significance of Data Lakes, especially in the context of modern data architectures, and underscores their potential impact on small-medium enterprises (SMEs) and startups. Data, often termed the lifeblood of any organization, plays a crucial role in modern business intelligence systems. Efficient and optimal data analytics provide a competitive edge, enhancing performance and services. As organizations grapple with the challenges of managing and analyzing the sheer volume and variety of big data, two data management systems have become prominent: data warehouses and data lakes (Nambiar & Mundra, 2022) [18]. While both serve as platforms to accumulate the vast data generated by organizations, they differ significantly in their characteristics and applications.

Data Lakes, as conceptualized by Pentaho CEO Jame Dixon, are akin to vast reservoirs storing data in its rawest form, unfiltered and unprocessed (Dixon, as cited in Khine & Wang, 2018) [23]. This contrasts with traditional data warehouses that store cleaned, processed data ready for consumption. The allure of Data Lakes lies in their ability to store every piece of data produced by an organization, offering insights at a much finer granularity (Miloslavskaya & Tolstoy, 2016) [17]. This is particularly relevant in the Big Data era, where traditional data storage and processing methods are being challenged by the sheer volume, velocity, and variety of data being generated. The inception of Data Lakes was driven by the business sector rather than academia, marking a shift in the way data storage concepts are typically developed (Miloslavskaya & Tolstoy, 2016) [17]. The fundamental premise of a Data Lake is straightforward: all data generated by an organization, irrespective of its format or structure, is stored in a singular structure known as the Data Lake. This eliminates the complex preprocessing and transformation typically associated with loading data into data warehouses, thereby reducing upfront data ingestion costs. From an architectural perspective, Data Lakes employ a flat architecture, storing data in its raw format. Each data entity within the lake is associated with a unique identifier and an extensive set of metadata (Nambiar & Mundra, 2022) [18]. This approach contrasts with traditional data warehouses, which are built on well-defined, structured schemas. Data Lakes, on the other hand, leverage a schema-on-read approach, where the data's nature and structure are determined at the time of querying.

For SMEs, startups, and innovation hubs like Isola, the adaptability and scalability of Data Lakes can be instrumental. As these entities work on cutting-edge projects that generate diverse types of data, the ability to store and process this data efficiently can significantly accelerate innovation cycles. Moreover, the flexibility offered by Data Lakes in terms of data storage and processing can be a game-changer for smaller businesses that may not have the resources to invest in traditional data warehousing solutions. However, the adoption of Data Lakes is not without challenges. The lack of a predefined schema means that businesses need to invest in robust metadata management strategies to ensure that data can be effectively queried and analyzed (Nambiar & Mundra, 2022) [18]. The sheer volume of data stored in Data Lakes can make data governance a complex endeavor. Hence, Data Lakes represent a paradigm shift in the way modern businesses approach data storage and analytics. Their ability to store vast amounts of raw data offers unparalleled flexibility and granularity in data analysis, making them an attractive proposition for SMEs, startups, and innovation hubs. As businesses continue to generate and rely on vast amounts of data, the role of Data Lakes in modern data architectures will only become more prominent.

2.2.1 Why AWS as Data Lake service

AWS has rigorously developed a data-lake architecture that utilises the capacity of Amazon Simple Storage Service (S3) and numerous supporting facilities. These services provide a range of specialised functionalities, including effortless amalgamation with conventional big data tools and pioneering query-in-place analytical tools. These tools effectively curtail expenses and intricacies by removing procedures such as data extraction, transformation, and load (Hukkeri et al., 2020) [10]. In addition, AWS's solution for data lakes integrates with services such as Amazon Kinesis Firehose, which offer crucial transformation functions, including data batching, compression, lambda functions, and encryption.

An exceptional attribute of Amazon S3 is its ability to version buckets, which establishes measures to avert data and project loss. This characteristic also ensures the optimal separation of concerns between various stages of the analytics project. Furthermore, Amazon S3 provides safeguarding measures that protect against personal data losses. It should be noted though that AWS's platform could potentially have slower performance when compared to that of the traditional Hadoop Distributed File System (HDFS). Despite this, the favourable characteristics of AWS's Data Lake, including its strong security design, flexibility, and high data durability, make it an attractive option for medium-small enterprises seeking to delve into the domain of Big Data.

The data architecture employed by AWS centralises data ingestion from various sources into one platform that facilitates easy integration with existing and future third-party data processing tools, making it a comprehensive, cost-effective, and efficient solution that is perfect for small to medium-sized enterprises. The data architecture employed by AWS centralises data ingestion from various sources into one platform that facilitates easy integration with existing and future third-party data processing tools, making it a comprehensive, cost-effective, and efficient solution that is perfect for small to medium-sized enterprises. This makes AWS one of the leading providers of data lake solutions. The data architecture employed by AWS centralises data ingestion from various sources into one platform that facilitates easy integration with existing and future third-party data processing tools, making it a comprehensive, cost-effective, and efficient solution that is perfect for small to medium-sized enterprises.

2.2.2 The Medallion Architecture

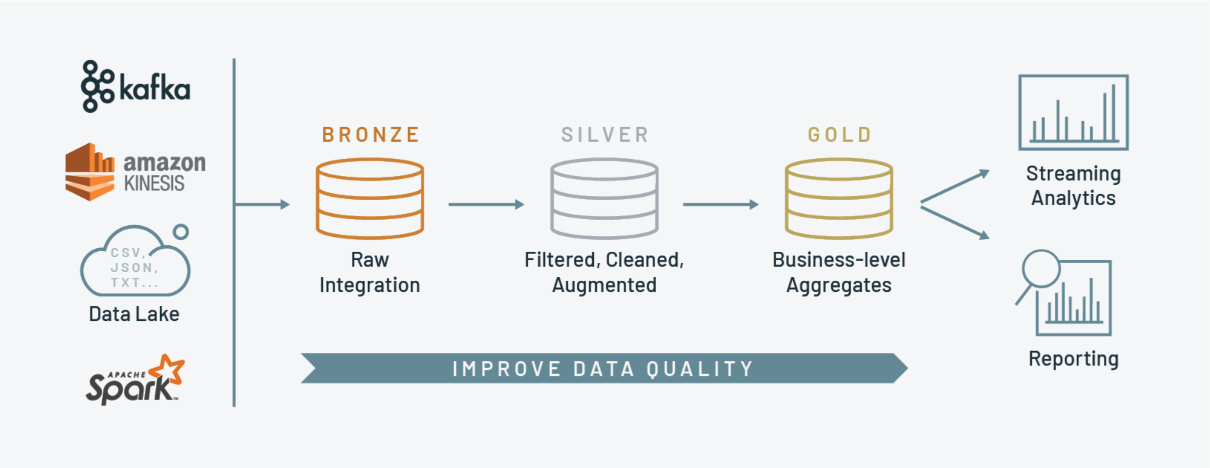


Figure 2.1: Medallion Architecture

The Medallion Architecture, also known as the *Bronze-Silver-Gold* pattern, is an influential framework in data management, particularly within the context of Data Lakes (L'Esteve, 2023) [15]. This approach, especially prominent in cloud-based data platforms like Databricks, AWS, and Azure, is structured to streamline the processing of large-scale data, maintain data integrity, and optimize the use of data for diverse analytics and machine learning applications. Here follows the main features of this storage architecture:

- **Bronze Layer:** The Raw Data Layer
  - **Foundation of Data Lake:** The Bronze layer serves as the foundational tier, where raw data is ingested from various sources.
  - **Nature of Data:** It houses unmodified, unprocessed data in its original format. This could include logs, IoT device data, files, database records, etc.
  - **Storage Characteristics:** The focus here is on reliable, fault-tolerant storage, making it a comprehensive repository or a single source of truth for all raw data.
  - **Data Types:** It encompasses a wide range of data types – structured, semi-structured, and unstructured.
  - **Importance for Data Recovery:** In case of any processing error in subsequent layers, the Bronze layer allows for data recovery and reprocessing.
- **Silver Layer:** The Refine Data Layer
  - **Transformation and Cleansing:** This layer is where raw data from the Bronze layer undergoes cleaning, transformation, and enrichment.
  - **Preparation for Use Cases:** It involves schema enforcement, deduplication, normalization, and error correction, refining the data for specific use cases.
  - **Data Enrichment:** The Silver layer may also involve the augmentation of raw data with additional context or insights to enhance its value.
  - **Analytics Readiness:** Here, the data starts becoming more structured and suitable for analysis but might not be in the final form for business reporting.
- **Gold Layer:** The Curated Data Layer
  - **Business-Ready Data:** The Gold layer contains data that has been highly curated and is optimized for direct business consumption.
  - **Data for Decision Making:** It typically includes data aggregated, summarized, or transformed to support decision-making processes, reporting, dashboards, and business intelligence.

- **End-User Focused:** This layer offers datasets tailored for end-user analysis, ensuring ease of understanding and interpretation.
- **Machine Learning and Advanced Analytics:** It is often used for more advanced purposes, including predictive analytics and machine learning models.

2.3 Cluster analysis

Clustering analysis is a quintessential facet of unsupervised learning, a domain of machine learning wherein the algorithm learns from unlabeled data to discover inherent structures within. It primarily deals with partitioning data into homogeneous groups or clusters, which share similar traits, without having prior labels for the groups. As stated by Xu & Tian (2015)[22], clustering analysis is to uncover hidden patterns in data, thereby providing valuable insights and aiding in decision-making across various fields. Over the years, a multitude of clustering algorithms have been devised to cater to different types of data and clustering requirements. Among the plethora of clustering algorithms, some have stood out owing to their efficacy, robustness, and wide applicability. These include:

- **Partitioning Algorithms:** K-means is a prime example, which seeks to partition data into K distinct, non-overlapping subsets (or clusters) based on their mean. The objective function J aims to minimize the within-cluster sum of squares (WCSS). Here, K represents the number of clusters, C_i represents the i-th cluster, x is a data point, and μ_i is the mean of the data points in cluster C_i. The formula essentially computes the squared distance of each data point from the centroid of its assigned cluster and sums these across all clusters.

$$J = \sum_{i=1}^{K} \sum_{x \in C_i} \|x - \mu_i\|^2 \tag{2.1}$$

- **Hierarchical Algorithms:** Agglomerative Nesting (AGNES) is a hierarchical method that builds a tree of clusters by successively merging or splitting existing groups. The distance $d(C_i, C_j)$ between two clusters C_i and C_j is computed using the single-linkage criterion, which considers the shortest distance between points in different clusters. Here, x and y are data points in clusters C_i and C_j, respectively.

$$d(C_i, C_j) = \min_{x \in C_i, y \in C_j} \|x - y\| \tag{2.2}$$

- **Density-based Algorithms:** DBSCAN (Density-Based Spatial Clustering of Applications with Noise) operates by growing clusters from regions of the data space where the density of data points is high. A point x is classified as a core point if the number of points within a specified radius ϵ is at least MinPts. $N_\epsilon(x)$ denotes the ϵ-neighborhood of point x.

$$N_\epsilon(x) \geq \text{MinPts} \tag{2.3}$$

- **Grid-based Algorithms:** STING (Statistical Information Grid) divides the data space into a finite number of cells, forming a grid structure, and then performs clustering on the grid. In STING, the data space is divided into a grid with a finite number of cells. Clustering is performed based on statistical information stored for each cell. There isn't a specific formula as the clustering is driven by the grid structure.

- **Soft Clustering Algorithms:** FANNY allows data points to belong to multiple clusters with different degrees of membership. The objective function J minimizes a weighted sum of squared errors, where w_{ix} is the degree of membership of point x in cluster C_i, and μ_i is the centroid of cluster C_i.

$$J = \sum_{i=1}^{K} \sum_{x \in C_i} w_{ix} \left\| x - \mu_i \right\|^2 \tag{2.4}$$

- **Self-Organizing Maps (SOM):** SOM use a neural network model to achieve clustering. In SOM, the weight vector $w(t)$ is updated iteratively to match the input data x. Here, $\alpha(t)$ is the learning rate, and $h_{ij}(t)$ is the neighborhood function that adjusts the learning rate based on the distance in the network topology.

$$w(t+1) = w(t) + \alpha(t) \cdot h_{ij}(t) \cdot (x - w(t)) \tag{2.5}$$

- **Ensemble Clustering:** This method combines multiple clustering solutions into a single consolidated clustering. The co-association matrix C averages the co-occurrences of pairs of data points i and j in the same cluster across multiple clustering solutions M. $C_k(i)$ and $C_k(j)$ represent the clusters of point i and j in the k-th clustering, and δ is the Kronecker delta function, which is 1 if the arguments are equal and 0 otherwise.

$$C(i,j) = \frac{1}{M}\sum_{k=1}^{M}\delta(C_k(i), C_k(j)) \tag{2.6}$$

The breadth of applications for clustering analysis is immense, with its relevance accentuated by the burgeoning quantity of data across diverse domains. Clustering facilitates the extraction of meaningful patterns from bulk data, a process integral to knowledge discovery. This is especially pertinent in the modern era, where data-driven decisions are paramount across myriad fields including but not limited to marketing, healthcare, finance, and many more. The ability to distill large datasets into actionable insights through clustering analysis is invaluable, making it an indispensable tool in the repertoire of data scientists and analysts (Ezugwu et al., 2022) [5].

2.3.1 Handling mixed data-types in cluster analysis

Dealing with mixed data types, which include both categorical (qualitative) and numerical (quantitative) variables, can be challenging in clustering analysis. This mixed nature of data poses unique challenges, as traditional clustering algorithms like k-means are typically not designed to handle such diversity in data types. As stated in Ghosal et al.(2019), categorical variables consist of values that represent different categories [7]. For our discussion, nominal, ordinal, and dichotomous variables are collectively referred to as categorical.

1. **Dichotomous Variables:** These have only two categories (e.g., yes/no, male/female) and are often coded as zero and one. The significance of each category varies; some are symmetric (both categories equally important), while others are asymmetric (one category more significant than the other).

2. **Multi-Categorical Variables:** These include:

   - **Nominal Variables:** Categories with no order or intensity (e.g., colors, types of cuisine).
   - **Ordinal Variables:** Categories with a specific order but arithmetic operations are not applicable (e.g., rankings, education levels).
   - **Quantitative Variables:** These allow arithmetic operations (e.g., age, income).

Handling mixed data, comprising both categorical and numerical variables, requires specialized techniques. The primary challenge lies in the inherent dissimilarity of data types. To address this, various methodologies have been developed:

1. **Preprocessing Techniques:**
    - **Discretization and Dummy-Coding:** These methods convert all variables into a single type (continuous or categorical), enabling the application of classical clustering techniques on a uniform dataset. Discretization bins continuous variables into categories, while dummy-coding creates binary columns for categorical variables. However, these methods may alter the original data's structure, potentially introducing bias.
2. **Specialized Clustering Algorithms:**
    - **k-Prototypes Algorithm:** This extends the k-means paradigm for mixed data by defining a combined dissimilarity measure and replacing cluster means with modes for categorical data.
    - **k-CMM:** Designed for mixed data with missing values, this algorithm combines imputation and clustering steps.
    - **Model-Based Clustering (MBC):** Assumes the data is generated from a mixture of underlying probability distributions, suitable for mixed data as it can assume different distributions for continuous and categorical variables.
    - **Algorithms Utilizing Dissimilarity Matrices:** Techniques like Partitioning Around Medoids (PAM) and Hierarchical Clustering can be adapted to mixed data using measures like the Gower distance that handle different data types.

In the paper *Cluster analysis and categorical data*, Hana Řezanková (2009) [20] explores various approaches to clustering in the context of categorical data. She highlights that while methods for cluster analysis of quantitative data are widely implemented, the methods for clustering qualitative data vary significantly. The paper offers a comprehensive understanding of clustering techniques and the importance of treating categorical data with precision and care. In conclusion, the intricate relationship between cluster analysis and categorical variables demands specialized techniques for accurate and meaningful results. This is especially true in the realm of mixed data, where the blend of categorical and numerical variables adds another layer of complexity, requiring thoughtful and tailored approaches in data analysis.

2.3.2 The k-Prototypes algorithm

The k-Prototypes algorithm, introduced by Huang in 1997, represents a significant enhancement to the traditional k-means clustering algorithm, specifically designed to address the complexities associated with mixed data types. Mixed data represent a challenge for standard clustering algorithms like k-means, which are primarily tailored for numerical data [9].

Huang's k-Prototypes algorithm integrates the principles of the k-means and k-modes algorithms, effectively combining their strengths to handle datasets with both numerical and categorical attributes. The k-means component of the algorithm is used for clustering numerical attributes, employing the conventional approach of minimizing the sum of squared distances within clusters. On the other hand, the k-modes component is applied to the categorical attributes, using modes instead of means for central tendency and a different dissimilarity measure appropriate for categorical data. One of the key innovations of the k-Prototypes algorithm is its cost function, which is a hybrid that considers both types of data attributes. The algorithm iteratively updates the centroids of the clusters for numerical data and the modes for categorical data. This iterative process continues until the assignment of data points to clusters stabilizes, indicating that the optimal clustering has been reached. The k-Prototypes algorithm has proved to be particularly useful in various real-world applications where data is inherently mixed, such as market segmentation, customer profiling, and bioinformatics. Its ability to seamlessly handle the dichotomy of data types makes it a powerful tool in the field of data mining and knowledge discovery, where mixed data is prevalent. Its innovative approach to combining numerical and categorical data clustering makes it a cornerstone algorithm in data mining and analysis tasks involving complex and diverse datasets [9].

- **Mathematical Formulation:**
  The objective function of the k-Prototypes algorithm is to minimize the sum of the dissimilarities between objects and the corresponding cluster prototypes. The dissimilarity measure d between an object x and a prototype p is defined as the sum of the dissimilarities for numerical and categorical attributes:

$$d(x,p) = \sum_{j=1}^{n}(x_j - p_j)^2 + \gamma \sum_{j=n+1}^{m} \delta(x_j, p_j) \tag{2.7}$$

where:

- n is the number of numerical attributes,

- m is the total number of attributes (numerical + categorical),
- γ is a weighting factor that balances the importance of numerical and categorical attributes,
- δ is a simple matching dissimilarity measure defined as:

$$\delta(a, b) = \begin{cases} 0 & \text{if } a = b \\ 1 & \text{if } a \neq b \end{cases} \tag{2.8}$$

The objective is to minimize the cost function J:

$$J = \sum_{i=1}^{K} \sum_{x \in C_i} d(x, p_i) \tag{2.9}$$

where:

- K is the number of clusters,
- C_i is the i-th cluster,
- p_i is the prototype of cluster C_i.

Algorithm Steps:

1. *Initialization*: Select K initial prototypes, each being a randomly selected object from the dataset.

2. *Assignment*: Assign each object to the nearest prototype using the dissimilarity measure d.

3. *Update*: Update the prototypes by computing the mean (for numerical attributes) and mode (for categorical attributes) of all objects in each cluster.

4. *Termination*: Repeat the assignment and update steps until the clusters no longer change or the change is below a specified threshold.

Advantages: The k-Prototypes algorithm provides a robust method for clustering mixed data, effectively bridging the gap between k-means and k-modes, making it a versatile choice in real-world scenarios where data often comprise a mix of numerical and categorical attributes.

Disadvantages: Similar to k-means, k-Prototypes requires the number of clusters K to be specified a priori, which may not be known in advance. The algorithm's performance can be sensitive to the initial placement of prototypes, and may converge to local minima, which is a common issue with iterative refinement clustering algorithms.

2.3.3 Hierachical clustering

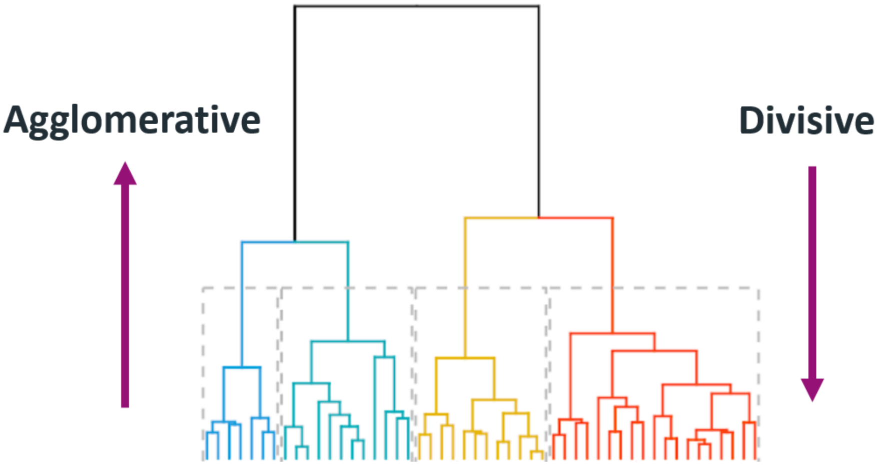


Figure 2.2: Example of Hierarchical Clustering

Hierarchical clustering is a method of cluster analysis that aims at building up a hierarchy of clusters, offering a unique approach compared to other clustering techniques. This method, which dates back to the foundational work by Johnson in 1967, constructs a tree-like structure called a dendrogram, illustrating how individual elements progressively merge into larger clusters [12]. At one end of the dendrogram are the individual elements, and at the other is a single cluster encompassing all elements. The process involves arranging objects so that those in the same cluster (or subgroup) are more similar to each other than to those in other clusters. Hierarchical clustering can be executed in two primary ways: Agglomerative and Divisive.

1. **Agglomerative (Bottom-Up):** This is the more common approach, where each element starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy. It's akin to a bottom-up approach where you start with many small pieces and combine them to form larger groups.

2. **Divisive (Top-Down):** In contrast, divisive clustering starts with all elements in one large cluster and progressively splits the cluster into smaller pieces. This method resembles a top-down approach, beginning with a large entity and progressively subdividing it into smaller segments.

The choice of distance measure and linkage method is crucial in hierarchical clustering, as they define the rules for grouping elements and clusters. These define how the similarity between two elements or clusters is calculated. Common measures include:

- **Euclidean Distance:** The Euclidean distance is the straight-line distance between two points in a multidimensional space.

  $$d(\mathbf{p}, \mathbf{q}) = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2} \tag{2.10}$$

  In this formula, $\mathbf{p}$ and $\mathbf{q}$ are two points in n-dimensional space, with p_i and q_i representing their coordinates in the i-th dimension.

- **Manhattan Distance:** The Manhattan distance computes the sum of the absolute differences of the coordinates of two points.

  $$d(\mathbf{p}, \mathbf{q}) = \sum_{i=1}^{n} |p_i - q_i| \tag{2.11}$$

  It's equivalent to the total distance traveled along axes at right angles in a grid-like path.

- **Cosine Distance** The Cosine distance measures the cosine of the angle between two vectors.

  $$d(\mathbf{p}, \mathbf{q}) = 1 - \frac{\mathbf{p} \cdot \mathbf{q}}{||\mathbf{p}||||\mathbf{q}||} \tag{2.12}$$

  This metric is particularly useful in high-dimensional spaces such as text mining and document clustering, where it indicates the similarity in orientation, irrespective of magnitude.

- **Gower Distance** The Gower distance is a versatile metric used for datasets with mixed types of data.

  $$d(\mathbf{p}, \mathbf{q}) = \frac{1}{n} \sum_{i=1}^{n} \frac{|p_i - q_i|}{R_i} \tag{2.13}$$

  Here, R_i is the range of the i-th variable, and n is the number of variables. This metric computes similarities based on the nature of the variables.

Linkage methods determine how the distances between clusters are computed. The most common are:

- **Single Linkage:** The distance between two clusters is defined as the shortest distance between two points in each cluster.
- **Complete Linkage:** The distance is the longest distance between two points in each cluster.
- **Average Linkage:** The average distance between each point in one cluster to every point in the other cluster.
- **Ward's Method:** Minimizes the total within-cluster variance, aiming to choose the pair of clusters that leads to a minimum increase in total within-cluster variance after merging.

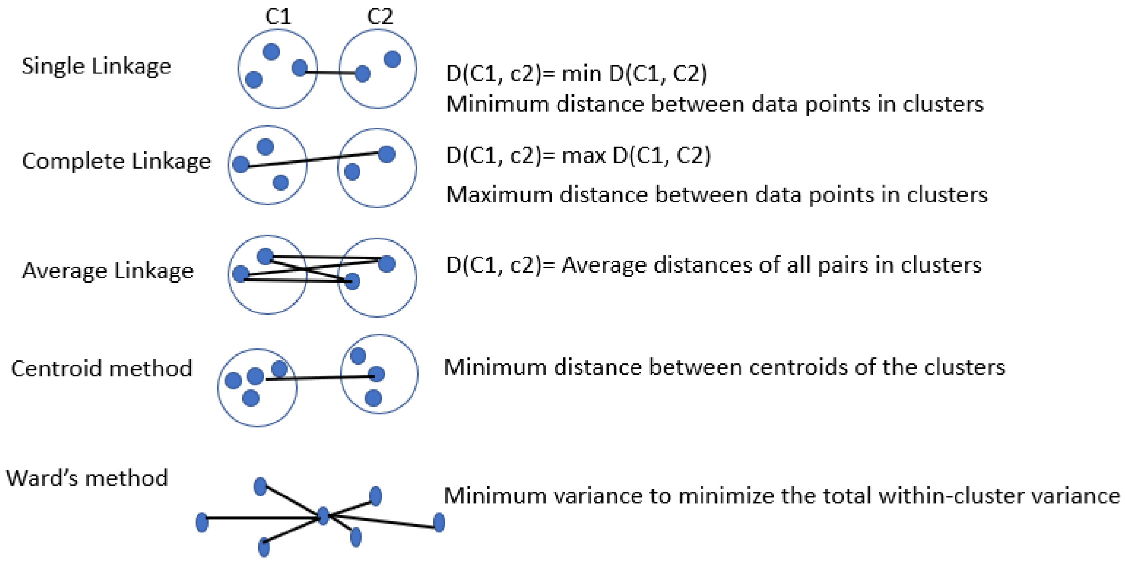

C1
C2
Single Linkage
D(C1, c2)= min D(C1, C2)
Minimum distance between data points in clusters
Complete Linkage
D(C1, c2)= max D(C1, C2)
Maximum distance between data points in clusters
Average Linkage
D(C1, c2)= Average distances of all pairs in clusters
Centroid method
Minimum distance between centroids of the clusters
Ward's method
Minimum variance to minimize the total within-cluster variance

Figure 2.3: Overview of different methods of linkage

2.3.4 Cluster analysis and its applications in user segmentation

Cluster analysis has been a cornerstone in the realm of data classification and interpretation. This technique's versatility has made it indispensable in various domains, especially in the intricate world of market segmentation and user profiling (Aldenderfer & Blashfield, 1984 [1]). Market segmentation, a pivotal facet of modern marketing, revolves around the division of a broad market into discernible segments that exhibit internal homogeneity. This means that individuals within a segment share certain characteristics, behaviors, or preferences, making them distinct from those in other segments (Runiewicz-Wardyn,

2014) [21]. This strategic division is not merely an academic exercise; it has proven its mettle in the real world by aiding managers in crafting marketing strategies tailored to specific audience subsets. The concept of market segmentation, though nearly six decades old, has evolved over time, with preference-based segmentation emerging as a contemporary and significant perspective. It's worth noting that variations in consumer preferences are foundational to market segmentation (Kardes, 1999) [14]. Such segmentation can be adeptly achieved using a combination of conjoint and cluster analysis, offering a nuanced view of the market landscape, ranging from simulating real-life purchasing scenarios to discerning the relative importance of product attributes (Punj et al. 1983) [19].

For innovation hubs and coworking spaces like Isola, comprehending the intricacies of its diverse userbase is of paramount importance. By harnessing the power of cluster analysis and its applications in user segmentation, Isola can glean invaluable insights into the preferences, needs, and behaviors of its users. This granular understanding can be the bedrock upon which tailored services, amenities, and community-building initiatives are built, thereby enhancing user satisfaction, fostering loyalty, and nurturing a sense of belonging. Moreover, coworking spaces, which often serve as incubators for startups and entrepreneurs, stand to gain immensely from such insights. By understanding the unique needs and preferences of these nascent businesses, coworking spaces can curate specialized programs, workshops, or partnerships that catalyze innovation, facilitate collaboration, and engender a sense of community. Speaking of which, the paper by Punj et al. (1983) [19] offers a comprehensive overview of the myriad benefits of preference-based market segmentation using conjoint and cluster analysis. They emphasize the technique's ability to craft questions that mirror real-life purchasing situations, minimize socially desirable responses, and provide a deep dive into the relative importance of product attributes. Such an approach, while traditionally associated with product-centric businesses, holds immense potential for spaces like Isola. By understanding and acting upon user preferences, coworking spaces can optimize space design, curate relevant community events, and offer services that resonate with their audience, thereby creating a vibrant, user-centric environment.

Methodology

3.1 The Data Sources

This section provides a brief overview of the main data sources and services that Isola uses to collect and store data:

Data source	**Platform**
Coworking Accesses	Microsoft Excel
Facility Check-in	Envoy
Paid reservations for coworking desks and meeting rooms	Office RnD
Newsletter and Marketing Email	Hubspot
Paid services and Fiscal data	Fatture in Cloud
Badge Access	Biticino
Ticket distribution and events check-in	Eventbrite
Analysis of island visibility and effectiveness of the first touchpoint	Google Analytics
User interactions	Instagram
User interactions	Youtube
User interactions	Facebook
User interactions	Linkedin
Wi-fi Network	Cambium Networks
Staff Email	Gmail

Table 3.1: Isola's main data source

- **Microsoft Excel**: Isola's coworking space uses Microsoft Excel to manage access logs and store information about coworkers and organizations. They employs Excel's data analysis and offline functionalities, such as pivot tables and formulas, to sort and examine user data.
- **Office RnD**: It is used to manage bookings at desks or meeting rooms within the facility. It includes people from local companies, legally and operationally associated,

those who bought meeting room hours or other membership options, individuals who self-registered for public facilities like drop-in coworking, and especially committed coworkers who frequently engage with Isola.

- **Hubspot**: It is used to supervise Isola's large network of contacts, communication, and promotional activities. As leading Customer relationship management (CRM) platform, it lets companies generate focused newsletters, manage contacts, and build communication strategies that actively involve their audiences. Additionally, the platform provides a personalised dataset specifically for delivering newsletters that keep members and potential customers informed.

- **Fatture in Cloud**: This financial management tool is fundamental for the hub's billing procedures, client relationship management, and comprehensive financial reporting. Fatture in Cloud provides the required tools and insights to issue invoices, track payments and analyze financial trends, as well as more. Through using Fatture in Cloud exports, the hub retains an elaborate client registry that is vital to both billing and relationship management.

- **Eventbrite**: Isola's community involvement strategy prioritises planning and promoting events. Eventbrite simplifies event management, covering registrations to feedback, for streamlined oversight over proceedings - be it a single workshop or a month-long series. This ensures members stay informed and involved. It's worth noting that data from specific months, like June 2023, carries added significance due to flagship events, such as "Make in South."

- **Envoy**: The Envoy platform simplifies access management by keeping track of all Isola check-ins and check-outs. This data provides valuable insights, including the rate of facility attendance and peak occupancy hours.

- **Google Analytics**: In today's digital world, having an online presence is essential. Isola utilises Google Analytics to measure its digital reach, providing invaluable insights into online searches surrounding the hub, as well as the geographical distribution of its visitors and the overall effectiveness of its digital strategies. Additionally, Isola has a digital footprint on various social media platforms, including Instagram, YouTube, Facebook, and LinkedIn. These platforms go beyond the role of simple content-sharing channels. They serve as communities in which Isola connects with its audience, shares its stories and cements its brand.

- **Cambium Networks**: Efficient digital experience is crucial within the hub. Recognising the importance of reliable and robust Wi-Fi connectivity in a coworking space, Isola adopts Cambium Networks to manage network usage, maintain access point health, and provide uninterrupted internet access to its members.

3.2 The Datasets

For this study, we are concentrating on three main areas. Firstly, we are examining the occupants of the shared working space, investigating their identity, behaviours, and group composition. Secondly, we are investigating meeting room reservations made by partner firms. Finally, we are evaluating the Isola-run events and the participant profiles at these gatherings. For each dataset analysed in this research, we provide a clear outline of the logical structure, variables, and primary cleaning and pre-processing steps.

3.2.1 Coworking Dataset

The dataset was sourced from the coworking space of Isola. It was generated over a period of almost two years, capturing real-time access records of individuals using the facility. Each entry represents a unique access instance, encompassing details about the user, their organizational affiliation, and the nature of their access. This dataset is derived from a larger database consisting of several tables. The main tables are *accessi_globale*, *anagrafica_utenti* and *anagrafica_organizzazioni*. These tables are in communication with each other thanks to a correspondence of the primary keys `id accesso`, `id utente` and `id organizzazione`.

Variable	**Description**
`nome`	Name of the organization
`id_organizzazione`	Unique identifier for the organization
`relazione`	Type of relationship with the coworking space
`profit`	Profit status of the organization
`nazione`	Country where the organization is based
`citta`	City where the organization is located
`indirizzo`	Address of the organization
`codice_fscale`	Fiscal code of the organization
`partita_iva`	VAT number of the organization
`pec`	Certified email address of the organization
`codice_univoco`	Unique code associated with the organization
`settore`	Business sector of the organization
`note`	Additional notes or comments about the organization

Table 3.2: *anagrafica_organizzazioni* dataframe

Variable	**Description**
`nome`	Name of the person
`cognome`	Surname of the person
`email`	Email address of the person
`id_persona`	Unique identifier for the person
`data_di_nascita`	Birthdate of the person
`sesso`	Gender of the person
`digital_nomad`	Indicator if the person is a digital nomad
`studente`	Indicator if the person is a student
`south_worker`	Indicator if the person is a south worker
`locale`	Indicator if the person is a local
`ospite`	Indicator if the person is a guest
`smart_worker`	Indicator if the person is a remote worker
`startupper`	Indicator if the person is a startupper
`posizione_lavorativa`	Job position
`settore`	Job sector
`telefono`	Telephone number
`nazione_cittadinanza`	Citizenship
`nazione_residenza`	Nation of residence
`note`	Further info

Table 3.3: *anagrafica_utenti* dataframe

Variable	**Description**
`id accesso`	Unique identifier for the access
`id utente`	Unique identifier for the user
`id organizzazione`	Unique identifier for the organization
`id pacchetto`	Package identifier
`data`	Date of access
`nome`	First name of the user
`cognome`	Last name of the user
`email`	Email address of the user
`stato`	Status of the access (e.g., paid, unpaid)
`tipologia`	Type of user (e.g., digital nomad, worker)
`settore`	User company sector
`azienda`	User company
`welcome email`	Welcome email status
`mail ringraziamento`	Thank-you email status
`note`	Additional notes about the user's access
`pagamento`	Fare paid by the user
`provenienza`	Country of origin

Table 3.4: *accessi_globale* dataframe

For our analysis we use the *accessi_globale* dataframe (Table: 3.4): The dataset is a complete and detailed record of every time an individual enters the coworking space. Each entry precisely documents different aspects of a user's experience in the space. Furthermore, we assume that each user can access the coworking space once a day, and that each user is associated with a single email address. Here is a detailed breakdown of the dataframe structure:

- **Identifiers:**
  - `id accesso`: This is a unique identifier for each access instance. It ensures that every entry in the dataset can be individually and distinctly referenced.
  - `id utente`: Representing each user with a unique identifier ensures that even if personal details like name or email change, the user's historical data remains interconnected. This ID is instrumental in user-centric analyses.
  - `id organizzazione`: For users affiliated with organizations, this identifier helps in grouping access instances by organizations.
- **User Personal Information:**
  - `nome` and `cognome`: The user's first and last names, respectively, provide straightforward identification. While these fields might not be unique, they aid in human-readable analyses and reports.
  - `email`: Capturing the email address facilitates communication with users. This field can also serve as an alternative identifier in certain scenarios.
- **Access-specific Details:**
  - `id pacchetto`: This field denotes the type or category of access the user has opted for. Different packages might offer varying amenities or time durations, and understanding their popularity can offer business insights.
  - `data`: A timestamp of when the access was recorded. This field is crucial for time series analyses, identifying peak periods, and understanding usage trends.
  - `stato`: This field captures the status of access, providing insights into whether a user has paid, if the access is complimentary, or if there's an outstanding payment. It can be instrumental in financial and operational analyses.
  - `pagamento`: This field indicates the fare that the user paid for accessing the coworking space, measured in euros. Analyzing this can provide insights into revenue patterns, user spending behaviors, and the popularity of different pricing tiers.

- **User Classification:**
  - `provenienza`: Captures the user's country of origin. This information can be vital for understanding the geographic diversity of the coworking space users and tailoring services or amenities to cater to specific cultural or regional preferences.
  - `tipologia`: This field classifies users into various categories. For instance, a digital nomad might use the space differently compared to a guest user. Understanding these patterns can help in tailoring amenities and services to specific user types.

    Below are the definitions of user types according to Isola:

    - *Digital Nomad*: A Digital Nomad is an individual who embraces a mobile lifestyle, working remotely and traveling, thereby contributing economically to both Isola and Catania through their way of life.
    - *Smart Worker*: A Smart Worker based in Catania chooses to live and work there, either as a freelancer or for local companies, enhancing Isola's economy without directly impacting Catania's economic landscape.
    - *South Worker*: A South Worker, originally from Catania, opts to work in the city for companies located outside Catania, Sicily, or even Italy, bringing economic advantages to both Isola and Catania through this local yet globally connected employment.
    - *Guest*: Guests are those who visit Isola occasionally or are invited by the staff, typically not contributing directly to the local work ecosystem but participating as visitors.

3.2.2 Eventbrite Dataset

Variable	**Description**
`id ordine`	Unique identifier for each order
`data ordine`	Date and time when the order was placed
`stato partecipante`	Status of the participant
`nome`	First name of the participant
`cognome`	Last name of the participant
`e-mail`	Email address of the participant
`nome evento`	Name or title of the event
`quantità biglietti`	Number of tickets ordered
`tipologia biglietto`	Type or category of the ticket
`prezzo biglietto`	Price of the ticket
`nome acquirente`	First name of the ticket buyer
`cognome acquirente`	Last name of the ticket buyer
`e-mail acquirente`	Email address of the ticket buyer
`currency`	Currency used for the transaction

Table 3.5: *Eventbrite* dataframe

Table 3.5 shows the *Eventbrite* dataframe which collects all events promoted or hosted within Isola. It provides a window into event attendance patterns, ticket purchasing behaviors, and other related dimensions:

- **Order Identifiers**: `id ordine` serves as a unique identifier for every order, allowing for distinct order tracking.

- **Timestamp**: `data ordine` specifies the exact date and time when a particular order was executed.

- **Participant Details**: Attributes such as `nome`, `cognome`, and `e-mail` capture the participant's first name, surname, and email address, respectively.

- **Event Information**: `nome evento` conveys the name of the event for which the order was placed, while `tipologia biglietto` and `prezzo biglietto` provide details on the type and price of the ticket purchased.

- **Purchaser Information**: While the participant details capture who will be attending the event, `nome acquirente`, `cognome acquirente`, and `e-mail acquirente` offer insights into the individual who made the purchase.

3.2.3 Office RnD dataset

Variable	**Description**
`company`	Name of the company affiliated with the user
`member`	Name of the individual member or user
`reference number`	Unique identifier for each booking or access
`start`	Timestamp marking the beginning of a reservation
`end`	Timestamp marking the end of a reservation
`resource`	Specific area or resource booked
`summary`	Brief description or title for the booking
`credits`	Internal currency or points used within the office space
`coins`	Another form of internal currency or points
`fee`	Cost associated with the booking, measured in euros
`extras`	Additional costs for supplementary services or amenities
`created at`	Timestamp for when the booking was made

Table 3.6: *Office RnD* dataframe

This *Office RnD* dataframe shown in Table 3.6 collects interactions between users and organisations that decide to rent desks or meeting rooms within the Isola facility. The booking process is fully managed by the Office RnD platform. Thanks to this system, the various users and companies can independently reserve and manage their bookings and block their time slots according to their own needs.

- **User and Company Information:**
  - `company`: Represents the company name to which the user is affiliated. This can reveal trends related to specific companies, such as their preferred resources or usage frequency.
  - `member`: The name of the individual member or user, providing clear identification for personal analyses.
  - `reference number`: A unique identifier for each booking or access, crucial for distinguishing and referencing individual entries.
- **Booking Details:**
  - `start` and `end`: Timestamps marking the beginning and end of a reservation. These fields are essential for analyzing usage patterns, identifying peak periods, and understanding how long resources are typically used.
  - `resource`: Indicates the specific area or resource booked, like a meeting room or a desk space. Analysis of this field can reveal the popularity and demand for various resources within the office space.

  - `summary`: Provides a brief description or title for the booking, offering insights into the purpose of reservations.

- **Financial Information:**
  - `fee`: The cost associated with the booking, measured in euros. This data is vital for financial analysis, understanding revenue streams, and user spending behavior.
  - `extras`: Additional costs incurred, possibly for supplementary services or amenities.

- **Miscellaneous:**
  - `created At`: The timestamp for when the booking was made, useful for tracking booking behaviors and forecasting future demands.
  - `credits` and `coins`: These fields might represent internal currency or points used within the office space, potentially indicating user engagement and loyalty programs.

3.3 Data Cleaning and Preprocessing

3.3.1 Missing data and data collection logic

The process of data cleansing and processing for all three datasets, especially for the *accessi_globale* dataframe, requires a meticulous and thorough approach, addressing various challenges inherent in the nature of data collection. The dataset heavily relies on manual data entry and has significant gaps, highlighting the complexities and limitations of the data collection process. During peak hours, or when users are reluctant to provide certain information, the dataset often suffers from missing data. This issue mainly occurs during user check-in, impacting significant personal details. It is particularly noticeable that key columns in the dataset, such as `tipologia`, `azienda`, `settore` and `provenienza`, have a significant lack of data, with up to 40% of values missing.

To address these data voids, a strategic approach to data imputation was implemented, complemented by a comprehensive reevaluation of the data collection methodologies. This effort is carried out in close collaboration with the Isola team, aiming to ensure that data is as complete and accurate as possible. The primary goal is to significantly diminish the gaps in the dataset, thus enhancing its reliability and utility.

In instances where direct imputation is not feasible, due to the nature or complexity of the missing data, placeholders are used. Terms like *undefined* and *unknown* are employed in such scenarios. These placeholders serve as temporary fill-ins, allowing for the continuity of data analysis while acknowledging the areas where data is incomplete or unavailable. This detailed effort in data cleaning and processing goes beyond mere technical fixes. It necessitates a comprehensive revision of the data collection methods, shifting from a potentially chaotic, traditional approach to a more precise, orderly, and data-focused strategy. This change will enhance the accuracy and reliability of the data, making it more valuable and understandable for both technical and non-technical audiences. For instance, in the `provenienza` column, we streamline the data to display only the user's country of origin, replacing earlier conflicting entries such as nationality, place of birth, or place of current residence.

As a further improvement, we introduce primary identification keys to facilitate join among tables. Values in the `tipologia` and `settore` columns are standardized. The `tipologia` column is restricted to four categories - smart workers, digital nomads, south workers, guests. Additionally, the `settore` column is refined to approximately ten distinct categories, accurately representing the companies in the Isola ecosystem. However, the processing for the other two datasets (i.e. *Eventbrite* and *Office RnD* dataframes) is less demanding since they are automatically generated, resulting in fewer cases of missing information, spelling errors or inconsistencies,

3.3.2 Variables selection

For these datasets, a preliminary analysis is conducted to ensure alignment with research goals and fulfil the varied information requirements of different levels of management and roles within Isola. Redundant columns, or those that present practical or privacy challenges to data collection, are removed. The variable selection process of this study follows two main principles:

1. First, it aims to eliminate columns that are redundant, unnecessary, or filled with missing values.

2. Second, it identifies fields most beneficial for exploratory analysis and capable of yielding significant insights.

These approaches vary across the different datasets: the *accessi_globale*, the *Eventbrite* and the *Office RnD* dataframes:

In *accessi_globale* (Table 3.4), this approach leads to the removal of less relevant columns such as `mail ringraziamento`, `welcome email` , `note`, and `numero`. These fields, while potentially informative in certain contexts, are not essential for the core analysis objectives.

For the *Eventbrite* dataset (Table 3.5), the focus centers on extracting meaningful data regarding event participation. Key columns for detailed analysis include those related to actual attendance, event dates, and titles. The `prezzo` category, in particular, offers valuable insights, showing that most of Isola's events are free, crucial for understanding the nature and accessibility of Isola's events.

The *Office RnD* dataset (Table 3.6) takes a different approach, emphasizing data about meeting room names and user information for both individual users and organizations. Columns detailing the start and end times of bookings are particularly important. From these, a new `durata` column is created to calculate the length of each booking in hours, providing clear insights into room utilization patterns and user preferences within the Isola ecosystem.

Overall, this careful process of variable selection and data refinement is key to tailoring each dataset for specific analytical purposes. It ensures the data not only meets immediate research needs but also lays a solid foundation for future exploratory analyses.

Building the Data Lake with AWS

In constructing the Data Lake for Isola, a suite of Amazon Web Services (AWS) has been strategically employed, each contributing unique functionalities essential for the efficient management and utilization of data. The core services involved in this process include Amazon S3, AWS Glue, AWS Lake Formation, and Amazon Athena. These services collectively create a robust, scalable, and sophisticated Data Lake architecture, well suited to meet Isola's diverse data requirements.

4.1 Amazon S3: The Storage Foundation

Figure 4.1: Amazon S3

Amazon S3, or Amazon Simple Storage Service, is essential to the data lake, providing a robust and secure object storage solution. With a durability of 99.999999999% and the scalability to handle large volumes of data, it is ideal for Isola, which collects data from many sources. S3's storage is distributed across multiple machines, significantly reducing the risk of data loss and outperforming on-premises solutions in terms of both durability and availability, although the latter varies depending on the class of storage chosen. The service can also handle different storage formats and is compatible with many AWS and third-party services, making it the perfect choice for a data lake that deals with many types of data and structures. S3 is highly scalable, handling from gigabytes to petabytes of data, which is ideal for Isola's growing data demands. Cost efficiency is also a considerable advantage: S3 has several storage classes, such as S3 Standard, Infrequent Access, and Glacier, all meant to offer budget-friendly solutions for diverse data access needs. This allows for cost-efficient scaling of storage without incurring unwarranted expenses. In addition to its storage capabilities, S3's advanced security features, such as integrated AWS security services, encryption (both in transit and at rest), access control policies, and logging capabilities, ensure enhanced data security. These aspects collectively make Amazon S3 a cost-effective, secure, and reliable foundation for Isola's Data Lake.

4.2 AWS Glue: ETL and Data Cataloging

Figure 4.2: Amazon Glue

AWS Glue, fully compatible with Amazon S3, is used for two additional Data Lake components: data processing and cataloguing. As regards data processing, AWS Glue comprises a fully managed extract, transform, and load (ETL) service based on Apache Spark, ideal for big data processing. It is a completely serverless service, and therefore eliminates the need for Isola to provision or manage resources, thereby reducing the complexity and total cost of operating a Data Lake. It also offers some basic features for creating and monitoring workflows of multiple jobs. It allows the automation of the extraction, transformation and loading operations. It supports a wide selection of data transformations, from simple mappings to complex data cleansing and enrichment tasks, tailored for different data types and structures. One can choose from 20 pre-built transformations using the NoCode approach or create custom code. All Spark transformations are accessible. Moreover, Isola can concentrate on analysing data instead of managing infrastructure since Glue automatically sets up the required resources and adjusts to the workload. The service also offers a data catalog feature that acts as a central metadata repository, making it easier to manage data assets over time.

4.3 AWS Lake Formation: Streamlining Data Lake Management

Figure 4.3: Amazon Lake Formation

AWS Lake Formation is a service specifically designed for Data Lake management. It simplifies the setup and ensures secure data access. With Lake Formation, tasks such as data ingestion, cataloging, and access control, which are traditionally complex and time-consuming, are significantly streamlined. Some features are shared with AWS Glue, however, while Glue does not offer specific tools for data lake management, Lake Formation specializes in this area, providing a user-friendly interface for defining security, governance, and auditing policies, thereby simplifying the management of data permissions across the Data Lake. This feature is particularly beneficial for Isola, ensuring that the right individuals have the necessary access to the data, in line with compliance and security requirements. It also integrates seamlessly with other AWS analytics and machine learning services, enhancing Isola's ability to extract meaningful insights and providing a unified platform for various data operations.

4.4 Amazon Athena: SQL Querying on Data Lake

Figure 4.4: Amazon Athena

Amazon Athena offers an interactive query service that enables SQL querying directly on data stored in S3 (as long as they are registered in a Glue database), making data analysis more accessible and efficient without the need for traditional data extraction and loading processes. Its serverless architecture ensures that Isola doesn't have to manage the infrastructure and can instead focus on the queries to run. Athena can handle complex queries on big data effectively. This allows users to explore structured and unstructured data interactively. AWS Athena's pricing model is based on the amount of data scanned per query and the cost is calculated per terabyte of data scanned. Athena does not charge for DDL (Data Definition Language) or failed queries and costs can be minimised by efficient query writing, e.g. by using partitioning and columnar formats such as Parquet or ORC.

Together, these AWS services provide a comprehensive, secure, and scalable architecture, ensuring efficient data storage, processing, and analysis capabilities. One can improve this Data Lake architecture in a number of ways, for example adding Amazon EMR for custom big data processing. Nevertheless, the combination of Amazon S3, AWS Glue, AWS Lake Formation, and Amazon Athena forms a comprehensive and integrated solution that represents an ideal balance Isola's Data Lake needs at the current time. Each service plays a specific role, from data storage and processing to management and analysis, creating an ecosystem that is both powerful and efficient. This ensures an infrastructure that is not only capable of handling current data needs but is also scalable and adaptable for future requirements.

4.5 Constructing a Data Lake: A Methodical Approach

This section aims to provide a clear and concise overview of Isola's Data Lake development process, showcasing the end-to-end integration of different AWS services and the logical progression from initial planning to final implementation.

4.5.1 Step one: The Design phase

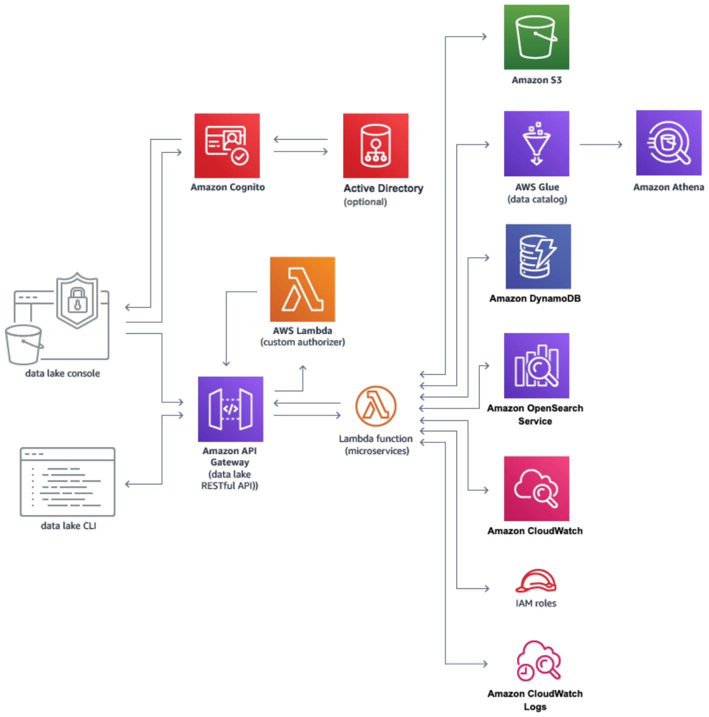

Amazon S3
Amazon Cognito
Active Directory (optional)
AWS Glue (data catalog)
Amazon Athena
Amazon DynamoDB
AWS Lambda (custom authorizer)
data lake console
Amazon OpenSearch Service
Amazon API Gateway (data lake RESTful API))
Lambda function (microservices)
data lake CLI
Amazon CloudWatch
IAM roles
Amazon CloudWatch Logs

Figure 4.5: Example of a Data Lake design using AWS services

The design phase lays the foundation for a robust, scalable, and secure data management system, essential to the organization's future data strategy. In this phase, all those steps that contribute to the overall efficiency and successful functioning of the Data Lake are conceived. It is important to ensure that the Data Lake is not only robust and secure, but above all that it is aligned with Isola's operational and analytical objectives. The initial phase of the project involves a detailed planning and design process:

1. It starts with defining Isola's objectives and scope for the data lake, analysing their short, medium and long term goals and determining the types of insights they want to extract. This analysis includes both current and anticipated future data requirements.

2. Having established the objectives, existing data sources are audited, either from internal sources (e.g. coworking space management or organised events) or from external sources such as social media analytics. In order to understand the diversity, volume and complexity of the data involved.

3. Data governance and compliance is then established to ensure data ownership, quality control, and adherence to regulations such as GDPR, embedding data privacy and security into the design from the start. In terms of security, permissions, and access control, the project employs AWS Data Lake Formation and IAM roles to establish granular access controls, defining who can view, modify, or delete data, and ensuring data integrity and confidentiality. Data protection is further reinforced through encryption of data both at rest and in transit, along with regular audits to ensure compliance with security standards.

4. Management and cost optimization are also crucial aspects of this phase that must not be overlooked: Isola can benefit from AWS cost management tools that allow monitoring and forecasting expenses related to data storage, processing and querying. For example, it is essential to efficiently plan scalable solutions to meet business growth while optimizing costs. Implementing data lifecycle policies and storage tiers are essential strategies that Isola can adopt for balancing data access needs with cost e.g. infrequently accessed data can be moved to cost-effective storage solutions such as Amazon S3 Glacier.

For Isola's Data Lake, a *medallion architecture* storage strategy (Section: 2.2.2), which incorporate a *Bronze*, *Silver*, and *Gold* layers architecture, was deemed as the most suitable option.

The Bronze Layer serves as a landing zone for Isola's core activities raw data in its native format, the Silver Layer is used for initial data cleansing and structuring, and the Gold Layer is reserved for fully processed and refined data, optimized for complex analytics and intended to be used by the management. In this initial phase of implementation, the project utilises three Amazon S3 buckets. One bucket is designated for storing raw and processed data, another holds ETL scripts, while a third contains queries saved from AWS Athena.

4.5.2 Step two: The ETL phase

The ETL phase is the second important step, allowing the raw data to be rationalised and transformed into actionable insights. This stage benefits from advanced AWS tools, particularly AWS Glue, to ease and automate the challenges of data integration.

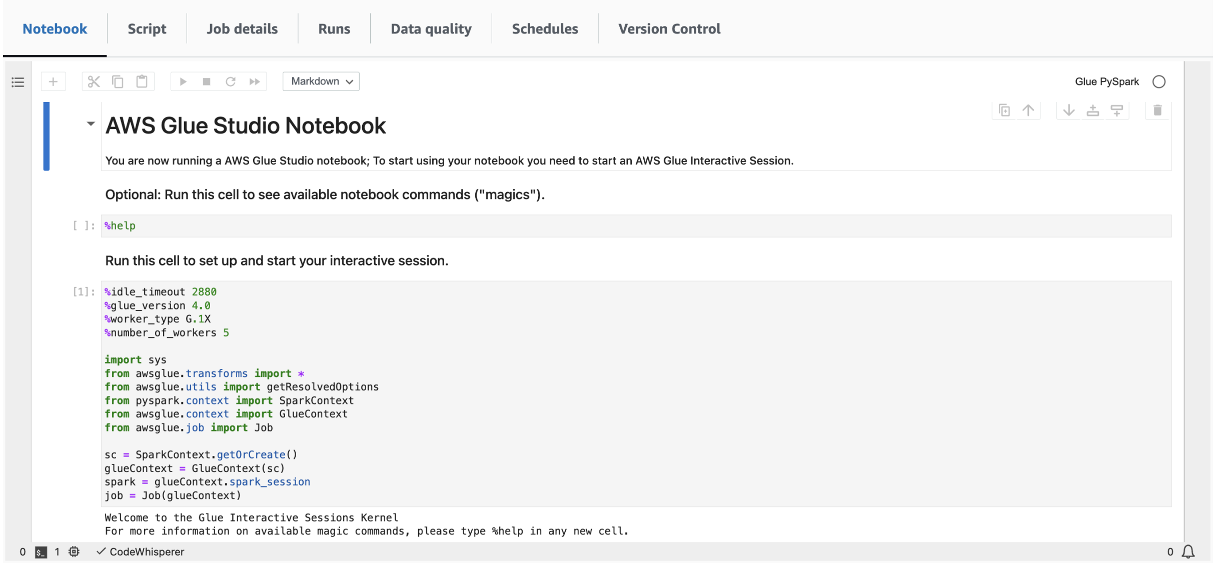
4.5.2.1 Extract

The process of ETL begins with data extraction, with AWS Glue central to automating the collection of data from various sources, both within and outside the organization. This stage is critical as it determines the effectiveness and efficiency of handling various data types. AWS Glue can be configured to effectively integrate with the diverse Isola's data sources, including company databases, CRM systems, cloud-based storage solutions, and external APIs. This integration is achieved by establishing data crawlers that effectively recognize and understand the format of the information available in these sources.

The initial step of data ingestion involves saving the collected data in its raw form in the *Bronze* layer of the Data Lake. This dataset comprises extensive information, incorporating data from coworking space management systems, social media feeds, event management platforms such as Eventbrite, and other relevant operational data for Isola.

4.5.2.2 Transform

The transformation phase is essential in the data processing journey as it converts raw data into a format suitable for analysis and decision-making. At this stage, inconsistencies, duplicates, and errors are eliminated through data cleansing to rectify discrepancies in data formats, correct spelling errors, and remove corrupt or irrelevant records. Data standardisation and normalisation are also essential in this stage. This process involves converting data into a consistent, comparable format to facilitate analysis. For example, it includes standardising date formats, unifying units of measurement, and normalising database schema. Additionally, data is augmented by adding new fields or integrating it with other dataframes. The process also involves organising and defining data schemas, aligning unstructured or semi-structured data into structured formats such as JSON or CSV.



AWS Glue Studio Notebook

You are now running a AWS Glue Studio notebook; To start using your notebook you need to start an AWS Glue Interactive Session.

Optional: Run this cell to see available notebook commands ("magics").

```
[ ]: %help
```

Run this cell to set up and start your interactive session.

```
[1]: %idle_timeout 2880
     %glue_version 4.0
     %worker_type G.1X
     %number_of_workers 5

     import sys
     from awsglue.transforms import *
     from awsglue.utils import getResolvedOptions
     from pyspark.context import SparkContext
     from awsglue.context import GlueContext
     from awsglue.job import Job

     sc = SparkContext.getOrCreate()
     glueContext = GlueContext(sc)
     spark = glueContext.spark_session
     job = Job(glueContext)
```

```
Welcome to the Glue Interactive Sessions Kernel
For more information on available magic commands, please type %help in any new cell.
```

Figure 4.6: Example of AWS Glue's interactive notebooks

AWS Glue offers a variety of data processing and cleaning tools, catering to different user needs and skill levels. Currently, there are several types of Glue ETL jobs available:

1. **Visual, Codeless Data Processing:** This user-friendly option allows for processing data visually without coding. It's particularly helpful for novices or non-technical users who prefer a more intuitive interface.

2. **Jupyter Notebook-Based Cleaning:** This method allows users to perform cleaning operations interactively using Jupyter notebooks. These notebooks run a version of PySpark tailored to the AWS Glue environment, offering a more hands-on approach to data manipulation (Figure 4.6).

3. **Script-Based Processing:** The more traditional and scalable method involves direct scripting. This approach requires a deeper level of technical knowledge, as users write and run scripts that process the input data in production. This method is ideal for those with programming experience and a need for higher performance and scalability.

For each dataset in this research, a dedicated notebook is created. These notebooks apply all necessary transformations and cleaning operations to address the data quality issues identified during the exploratory analysis, as documented in Chapter 3. This comprehensive approach ensures that the data is optimally prepared for further analysis and application.

- In *accessi_globale*, the cleaning process starts with an AWS Glue script that reads raw data and creates a dynamic frame for the initial inspection. This step includes removing irrelevant fields for the analysis, like `welcome email`, `note`, `mail ringraziamento`, among others. The data then transforms into a PySpark DataFrame, allowing more flexible manipulation and transformation. Essential operations involve casting the `data` column to a date format and adding temporal attributes such as month, year, and day of the week. Missing values in columns like `id pacchetto`, `email`, `tipologia`, `provenienza`,`settore`, and `stato` get filled with placeholders values to maintain structural consistency.

- *Eventbrite* dataframe cleansing process requires less effort compared to latter dataset. `Data ordine` variable is transformed into timestamp format, enabling temporal analysis. The `prezzo biglietto` field is converted to a float type, since it was coded as a string.
- The pre-processing of *Office RnD* was based to allow Isola to extract insights regarding the use of desks and meeting rooms. In this regard, some unnecessary fields were eliminated, and the focus was on the handling of time variables (e.g. room reservation date, reservation start time, reservation end time).

Once that data quality and cleaning is assured, the datasets are stored in the *Silver* layer of the Data Lake in CSV format.

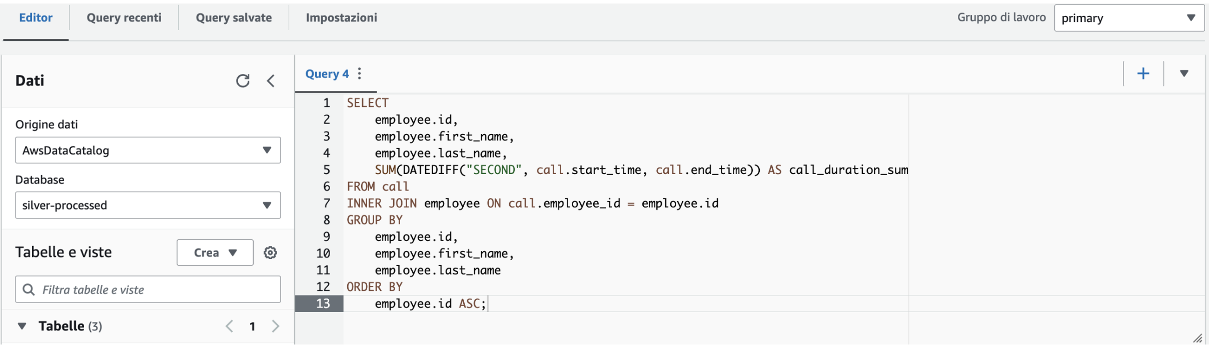
4.5.2.3 Load

In the final phase of the process, the data, once processed, is transferred and assigned to the *Silver* layer, which serves as an intermediary stage for the data. Despite being cleansed and transformed, it is not yet in its most refined state for analysis at this stage. This layer serves as a platform for initial analysis, preparing the data for further exploration and processing from the business. The data journey ends with its integration into the *Gold* layer, the ultimate repository in the Data Lake. Here, data undergoes final optimization stages, such as aggregation and indexing, perfectly preparing it for high-level analysis and business intelligence applications. The *Gold* layer is the ultimate in data refinement, hosting information that is tailored to meet the analytical needs of Isola's management and assist in a well-informed decision-making process.

4.5.3 Step three: The Investigation phase

The investigation phase employs AWS Athena to extract valuable insights. The service serves as a tool to execute complex SQL queries on the data stored in Amazon S3. Thanks to Athena, Isola can identify trends, patterns, and vital insights, particularly focusing on the behaviors and preferences of its user base. The analysis can range from examining peak usage times, user demographics, and interactions with the facility. The remarkable thing is that these insights can, in turn, improve the ETL processes that generated them. These can offer beneficial guidance to managers in improving their data manipulation. In fact, the ETL pipelines are designed to be modular and flexible to adapt easily to new evolving data requirements and trends. In the final stage, advanced reporting techniques are used to produce detailed insight reports. Those can be even generated through AWS visualisation tools or by integrating with third-party software such as Tableau or Power BI, offering Isola meaningful insights. This helps them make informed decisions and formulate strategies that align with the company's objectives.

4.6 Extracting knowledge through AWS Athena



Gruppo di lavoro: primary

Origine dati: AwsDataCatalog

Database: silver-processed

```sql
SELECT
    employee.id,
    employee.first_name,
    employee.last_name,
    SUM(DATEDIFF("SECOND", call.start_time, call.end_time)) AS call_duration_sum
FROM call
INNER JOIN employee ON call.employee_id = employee.id
GROUP BY
    employee.id,
    employee.first_name,
    employee.last_name
ORDER BY
    employee.id ASC;
```

Figure 4.7: AWS Athena console interface

The goal of this in-depth study of Isola's user base, is to permit Isola to identify usage patterns, demographic profiles and interactions within the facilities. This information provides invaluable insights about space utilization and highlights areas that could benefit from improvement or expansion. Also analysing feedback from events and workshops provides crucial information on the types of events that resonate with the community, engagement levels and participant feedback. This information is essential for evaluating the effectiveness of these activities and permit the management to plan future ones.

In the upcoming section, some detailed examples of business-ready SQL queries are presented. Isola's team can leverage them to extract valuable insights for analysis and decision making. The following queries show how a robust and reliable data infrastructure can enable a company to transform hidden and unexploited information into beneficial insights and actionable intelligence, driving informed decision-making and fostering a culture of data-driven growth.

4.6.1 Coworking Queries

Query 1: Count of Accesses and Revenues Per Month and Year This query helps in understanding the monthly trend of coworking space usage and revenues over different years.

```sql
SELECT ANNO,
     MESE,
     COUNT(email) as Total_Accesses,
     SUM(pagamento) as Total_Revenues
FROM processed_coworking
GROUP BY ANNO,
     MESE
ORDER BY ANNO,
     MESE;
```

Query 2: Accesses, Revenues and Number and Companies per Sector Understanding which sectors are most prevalent in the coworking space can guide business development and community-building efforts.

```sql
SELECT settore,
     COUNT(email) as Sector_Count,
     COUNT(DISTINCT(azienda)) as Companies_Count,
     SUM(pagamento) as Total_Revenues
FROM processed_coworking
WHERE settore IS NOT NULL
GROUP BY settore
ORDER BY Sector_Count DESC;
```

SQL Query 3: Distribution of User Types and Total Revenues per User Types Analyzes the distribution of different user types and computes the total revenues generated by each user type.

```sql
SELECT
    COALESCE(tipologia, 'undefined') AS tipologia,
    COUNT(email) AS Type_Count,
    SUM(pagamento) AS Total_Revenue,
    (COUNT(email) * 100.0 / (SELECT COUNT(email) FROM processed_coworking
        )) AS Percentage
FROM
    processed_coworking
GROUP BY
    COALESCE(tipologia, 'undefined')
ORDER BY
    Type_Count DESC;
```

4.6.2 Queries Eventbrite

Query 1: Count of Participants Per Event: This query groups the data by the event name and counts the number of participants for each event, ordering them by the count in descending order.

```
SELECT "nome evento",
     COUNT("e-mail") as Total_Participants
FROM processed_eventbrite
WHERE "stato partecipante" = 'Ha partecipato'
GROUP BY "nome evento"
ORDER BY Total_Participants DESC;
```

Query 2: Attendance ratio: This query calculates the monthly attendance ratio for events by dividing the number of actual attendees by the number of bookings, grouped and ordered by year and month.

```
SELECT
    EXTRACT(ANNO FROM "event_date") AS "Year",
    EXTRACT(MESE FROM "event_date") AS "Month",
    SUM(CASE WHEN "Stato partecipante" = 'Ha partecipato' THEN 1 ELSE 0
        END) AS "Ha_Partecipato_Count",
    SUM(CASE WHEN "Stato partecipante" = 'Parteciper ' THEN 1 ELSE 0 END
        ) AS "Parteciper_Count",
    CASE
        WHEN SUM(CASE WHEN "Stato partecipante" = 'Parteciper ' THEN 1
            ELSE 0 END) > 0
        THEN SUM(CASE WHEN "Stato partecipante" = 'Ha partecipato' THEN 1
             ELSE 0 END) * 1.0
             / SUM(CASE WHEN "Stato partecipante" = 'Parteciper ' THEN 1
                 ELSE 0 END)
        ELSE 0
    END AS "Attendance_Ratio"
FROM
    processed_eventbrite
GROUP BY
    "Year",
    "Month"
ORDER BY
    "Year" DESC,
    "Month" DESC
LIMIT 10;
```

Query 3: Monthly Event Count: This query counts the distinct events occurring each month, providing insight into the frequency of events over time.

```
SELECT ANNO,
       MESE,
       COUNT(DISTINCT "nome evento") as Event_Count,
       COUNT("e-mail") as Total_Participants
FROM processed_eventbrite
WHERE "stato partecipante" = 'Parteciper '
GROUP BY ANNO, MESE
ORDER BY ANNO, MESE;
```

4.6.3 Queries Office RnD

Query 1: Duration of Each Booking: This query calculates the duration of each booking by subtracting the start time from the end time. The TIMESTAMPDIFF function is used to find the difference in hours.

```
   SELECT Reference Number, Start, End,
          TIMESTAMPDIFF(HOUR, Start, End) AS Duration,
          Resource
   FROM bookings
   WHERE Resource IN ('Creta', 'Mallorca', 'Veglia', 'Gerba', 'Imbro')
   ORDER BY Start;
```

Query 2: Average Duration of Bookings Per Room: This query calculates the average duration of bookings for each of the specified meeting rooms.

```
   SELECT Resource, AVG(TIMESTAMPDIFF(HOUR, Start, End)) as
      Average_Duration
   FROM processed_officernd
   WHERE Resource IN ('Creta', 'Mallorca', 'Veglia', 'Gerba', 'Imbro')
   GROUP BY Resource;
```

Query 3: Total Hours Booked Per Company: This query sums the total hours booked by each company for the specified meeting rooms.

```
   SELECT Company,
   SUM(TIMESTAMPDIFF(HOUR, Start, End)) as Total_Hours
   COUNT("reference number") as Total_Bookings
   FROM processed_officernd
   WHERE Resource IN ('Creta', 'Mallorca', 'Veglia', 'Gerba', 'Imbro')
   GROUP BY Company
   ORDER BY Total_Hours DESC;
```

Cluster Analysis for User Segmentation

5.1 Pre-processing and feature engineering

Variable	**Description**
`sesso`	Gender of the user.
`tipologia`	Type or category of the user.
`settore`	Sector of the user's company
`avg_payment`	Average payment level of the user.
`std_payment`	Standard deviation of the user's payment.
`origin`	Origin of the user.
`total_accesses`	Indicates the volume of presence in the coworking space.

Table 5.1: Final variable selection for the cluster analysis

For this clustering analysis, the focus is on the *accessi_globale* dataframe with the purpose of segmenting and profiling users by identifying distinct clusters. This requires significant preprocessing to adapt the original dataset which logs individual instances of coworking access, to analyse the composition of the user base. A *group by* function is employed to aggregate individual accesses, transforming the access-level analysis into a user-level analysis. Table 5.1 presents the list of the selected variables, the aim is to create a complete profile for every user, incorporating a mix of old and novel developed features. These new variables originate from either aggregation (e.g. *average*, *sum*, *count* operations) or outcomes of feature engineering, which facilitates the extraction of new information from pre-existing data. The final selection is a result of an extensive revision of the *accessi_globale* dataframe, containing both original and newly produced variables. Moreover, we incorporate the *anagrafica_utenti* frame outlined in Table 3.3. Although the dataset is still subject to ongoing development, it already has the desired structure for our final dataframe, where each entry represents an individual user. This integration allows us to include extra information such as the users' gender, which is not available in the global access table. To achieve effective clustering, strategic decisions are made in designing the analysis:

- For the sole numeric variable `pagamento` in the dataset, two additional features are derived during aggregate analysis: `avg_payment` indicates the average user spending, and `std_payment` measures the standard deviation from this average. These features help in identifying users with higher spending tendencies, which may inform the development of bespoke packages or services.

- Adjustments are also made to two other variables to improve group uniformity and reduce variability within the group. The `provenienza` variable, initially documenting each user's nationality, is simplified to *italian* or *foreign* based on feedback from the Isola team. This change aligns with the business objective of differentiating between Italian and foreign users while minimizing extraneous data noise. Additionally, a new metric now quantifies users' total presence within the coworking space. This metric calculates the total number of accesses for each user and categorizes them into various levels based on their frequency of presence. These levels are determined by categorizing the total number of user presences into distinct groups, known as bins. The range of each bin is defined specifically: the first bin includes users with 1 to 4 presences, the second bin encompasses those with 5 to 9 presences, the third bin captures users with 10 to 19 presences, followed by a bin for 20 to 29 presences, and then 30 to an indefinite number of presences, allowing for the categorization of all users according to their frequency of visits.

5.2 Selection of clustering algorithms

5.2.1 Prelimanry considerations

The following analysis uses thus a diverse dataset containing quantitative and qualitative data. The varied mix of data types significantly impacts the selection of clustering algorithms, as already discussed in Chapter 2 of this research. The handling of categorical variables during clustering analysis is distinctively challenging. Since categorical variables are represented by words or labels, they are unsuitable for most machine learning algorithms developed for numerical data. As a result, these strings must be converted to a numerical format, which can be a complex process. The transformation technique chosen depends on the analyst's expertise, the dataset's context and the specific algorithm requirements. Moreover, numerical data typically undergo normalization or standardisation to ensure that they are adjusted to a common scale, thereby preventing any single variable from having a disproportionate influence on the analysis due to its scale.

In this study, no specialized transformations are applied to categorical variables during the preprocessing phase. This decision is based on the selected algorithms for analysis, namely Hierarchical Clustering and K-prototypes, since their implementation in Python language can automatically handle and transform these variables during the analysis. This feature makes the preprocessing stage simpler and allows for more direct analysis of the data in its original form. Conducting a clustering analysis, which combines art and science, necessitates a thorough comprehension of both the data and its domain. The intricacy of this process can be seen in algorithms like K-means or K-prototypes, where the number of clusters must be defined in advance. As clustering is unsupervised and requires significant domain knowledge, determining this number is not a trivial matter. To tackle these issues, the analysis starts with hierarchical clustering, which is appreciated for its exploratory nature in understanding the data structure. Throughout the analysis, we have tested several combinations of linkage methods, which play a crucial role in achieving effective clustering of the data by defining how the distance between data points is calculated. The objective is to determine the most suitable technique that corresponds to the statistical features of the data and the overall business objectives. To achieve this, we will be using the Gower distance metric (Equation: 2.3.3), which can handle various data types.

5.2.2 Hierarchical clustering

Python is selected as the programming language and Visual Studio Code as the development environment for conducting the Hierarchical Cluster Analysis (HCA). As previously noted, this HCA utilizes the Gower distance from the Python package *gower* and leverages the HCA implementation provided by the *scipy* package. The starting point of this analysis is the creation of a Gower distance matrix. This particular distance measure is selected for its ability to aptly handle a combination of numerical and categorical variables. Unlike traditional distance metrics like Euclidean or Manhattan, which are better suited for datasets comprising solely numerical data, the Gower distance metric can adeptly navigate the nuances of mixed data, providing a more accurate representation of the dissimilarities between data points.

After establishing the Gower distance matrix, the analysis actively explores various linkage methods commonly used in hierarchical clustering. These methods include single, complete, average, and Ward's linkage, each providing unique insights into how data points group together. The choice of method significantly influences the structure of the resulting data. Implementing a range of linkage methods demonstrates a holistic approach to uncovering the most intrinsic clusterings of data points, thereby enhancing the robustness and reliability of the clustering results.

The outputs of these linkage methods are visualized in the form of dendrograms. These tree-like diagrams are fundamental in hierarchical clustering as they provide a visual representation of the data points merging into clusters at various levels of similarity. Dendrograms are insightful as they not only show the formation of clusters but also the hierarchical relationships between them. The vertical axis of a dendrogram indicates the level of distance or dissimilarity at which clusters are merged, serving as a pivotal guide in deciding the optimal number of clusters. After examining all the dendrograms originating from the different linking methods, the most convincing output is the one in Fig 5.1, created using Ward as the linking method. In fact, this output gives us the clearest and most defined view of the clusters. Upon close visual inspection, three well-defined clusters are easily discernible.

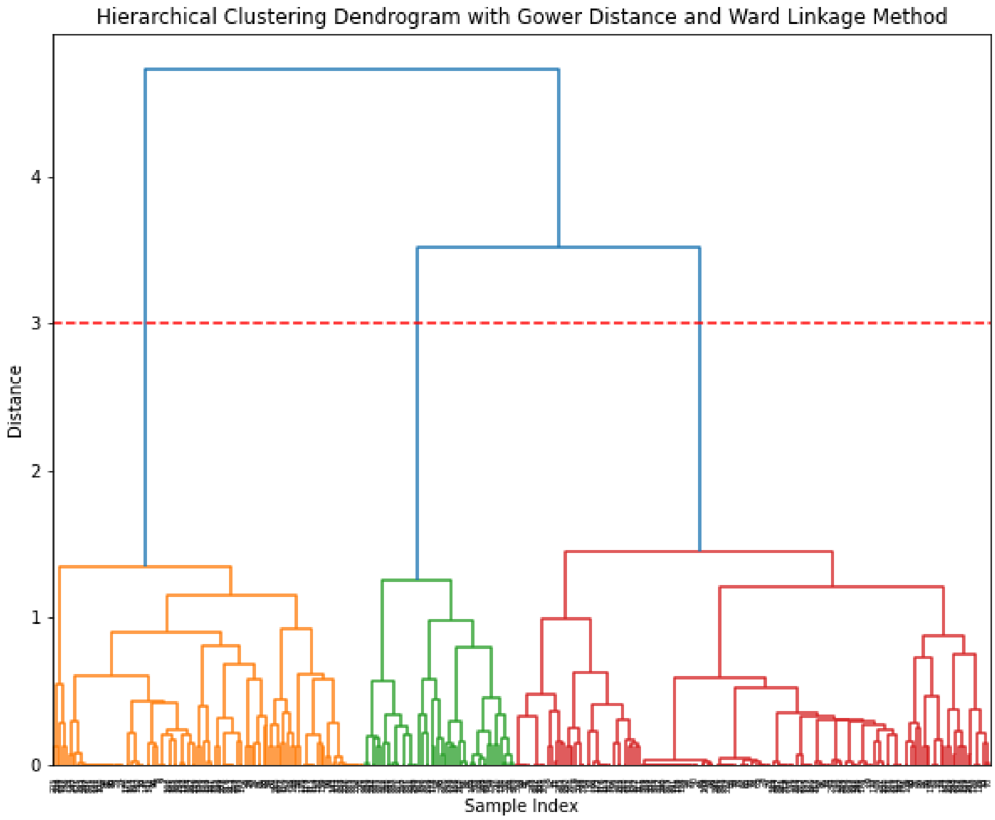


Figure 5.1: Dendrogram with Gower distance and Ward linkage method

Post dendrogram analysis, the identified clusters are then integrated back into the original dataset. This is a crucial step as it presents an opportunity to interpret the clusters in the context of the dataset's unique attributes. The integration process enables a detailed profiling of each cluster, allowing for the identification of key characteristics or variables that differentiate one cluster from another.

5.2.3 K-prototypes

Similarly HAC, the K-prototypes experiments are conducted using Visual Studio Code, with Python as programming language. The *kmodes.kprototypes* module is selected for its effectiveness as the algorithm implementation.

In this process, the *KPrototypes* object is instantiated with the number of clusters set to three. This choice reflects the results obtained by the HAC and suggested by domain expertise. Different initialization method are tested such as Cao or Huang. The choice of the Cao method for initialization indicates an intention to strategically position the initial cluster centroids, a step that significantly affects the efficiency and accuracy of the clustering results and ensures effective convergence to an optimal solution. Renowned for its effectiveness, this method determines an advantageous starting position for the cluster centroids, a crucial factor in K-prototypes clustering algorithms. This method employs a frequency-based approach to locate dense regions in the data space, positioning the initial centroids in areas representative of the underlying data distribution. This careful placement of centroids is critical for two main reasons. First, it enhances the efficiency of the clustering process. Starting from a position closer to the optimal, the algorithm needs fewer iterations to converge, speeding up the process. Second, and perhaps more crucially, it improves the accuracy of the clustering results. Effective initial positioning of the centroids increases the likelihood that the algorithm will discover the most natural and meaningful divisions within the data, avoiding misdirection by arbitrary or less representative starting points.

After configuring the K-prototypes model, it is applied to the dataset. A vital step in this application is the clear identification of which columns are categorical. This distinction is pivotal because it ensures the algorithm applies the most appropriate measure of dissimilarity for each type of data. The K-prototypes algorithm, by its design, ingeniously amalgamates the methodologies of K-means and K-modes clustering. This hybrid approach empowers the algorithm to adeptly process datasets that contain a mix of data types, a common scenario in real-world datasets. The output from this clustering process is the formation of distinct clusters, which are then merged back into the original dataset.

This merging is executed through the creation of a new dataframe that seamlessly integrates the original data with the cluster labels derived from the analysis.

Findings and Results

6.1 Centroid-based clustering

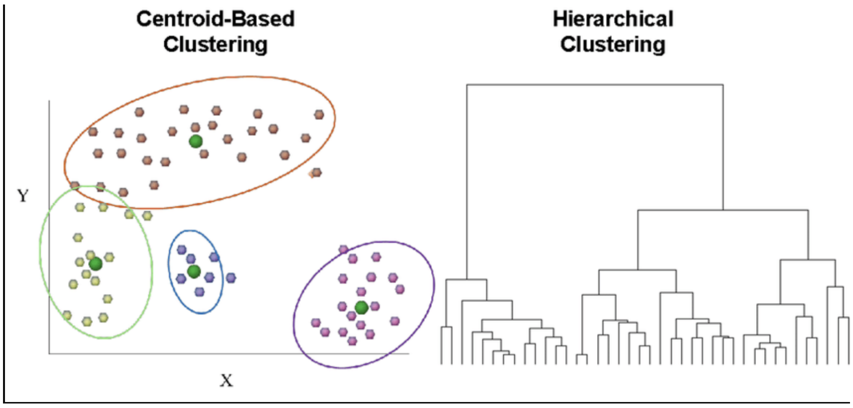


Figure 6.1: Illustration comparing centroid-based clustering and hierarchical clustering

Centroids, as a result of K-prototypes analysis, are essentially the central points of the clusters formed during the analysis. In K-prototypes, centroids play a crucial role. Each of them represents the average position of all data points in its cluster, effectively summarizing the cluster's overall characteristics. In the context of numerical attributes, a centroid's coordinates are the mean values of the attributes of all the data points in that cluster. For categorical attributes, the centroid is typically determined by the mode (the most frequent category) of the data points in the cluster. The process of K-prototypes clustering involves finding the best set of centroids such that the distances between the data points and their respective centroids are minimized. This ensures that each data point is as close as possible to the centroid of the cluster it belongs to, implying high intra-cluster similarity and distinctness from other clusters. In Figure 6.1, we can observe how centroid-based clustering and hierarchical clustering work differently. Evaluating the consistency of results across all methods used is an effective way to ensure data stability and algorithm robustness.

6.1.1 Centroid-Based user archetypes

Cluster	Cardinality
1	116
2	192
3	50

Table 6.1: Cardinality of clusters

The K-Prototypes analysis performed in Chapter 5 identifies three distinct centroids representing different user archetypes in the coworking space. Each centroid is characterized by specific attributes such as spending patterns, gender, user type, work sector, nationality, and access frequency.

- Centroid 1 - *The Modest Spenders*

  Table 6.2: Attributes of Centroid 1

Attribute	Value
Avg. Money Spent (Normalized)	0.035
Std. Dev. of Total Spent (Normalized)	0.030
Coefficient of Variation	85.71%
Gender	female
User Type	smart Worker
Work Sector	third Sector
Nationality	italian
Access Frequency	[1.0, 5.0)

  Centroid 1, labeled *The Modest Spenders*, is characterized by relatively moderate spending patterns, with a normalized[1] average money spent value of 0.0353. The standard deviation of total spent is 0.0302, indicating stable and consistent spending behavior. This cluster primarily consists of female users who are identified as smart workers in the third sector. They are predominantly Italian, showing a pattern of infrequent access to the coworking space, usually between one to five times.

- Centroid 2 - *The Minimalist Users*

  Centroid 2 represents *The Minimalist Users*, indicating a group with cautious financial behavior within the coworking space, as shown by a low normalized average

[1] A *normalized* variable is a variable that has been adjusted to a standard range, usually between 0 and 1. This allows for fair comparisons of diverse data sets while reducing the impact of differences in size or unit.

Table 6.3: Attributes of Centroid 2

Attribute	**Value**
Avg. Money Spent (Normalized)	0.052
Std. Dev. of Total Spent (Normalized)	0.029
Coefficient of Variation	55.77%
Gender	maschile
User Type	smart worker
Work Sector	third sector
Nationality	italian
Access Frequency	[1.0, 5.0)

money spent value of 0.0524. The standard deviation for this cluster is 0.0291, suggesting a very consistent and restrained spending pattern. This cluster comprises male smart workers engaged in the third sector, similar to Centroid 1 in terms of spending and access patterns, but different in gender.

- Centroid 3 - *The Frequent and High Spenders*

Table 6.4: Attributes of Centroid 3

Attribute	**Value**
Avg. Money Spent (Normalized)	0.777
Std. Dev. of Total Spent (Normalized)	0.931
Coefficient of Variation	119.7%
Gender	femminile
User Type	smart worker
Work Sector	digital & communication
Nationality	italian
Access Frequency	[20.0, 30.0)

Centroid 3 encompasses *The Frequent and High Spenders*, standing out with significantly higher spending levels. The normalized average money spent is 0.7774, and the standard deviation is notably high at 0.9311. This indicates not only a higher level of spending but also substantial variability in spending amounts. The cluster includes female smart workers primarily in the digital and communication sector, differentiating from the other centroids in both spending and visit patterns. They are characterized by Italian nationality and more frequent visits to the coworking space, typically between 20 to 30 times.

6.1.2 Cluster Distribution Analysis

The following subsections provide a more in-depth overview of the distribution of coworking space users' attributes across three identified clusters, with accompanying visualizations. As shown in table 6.1, the cardinality of the three clusters is not consistent. To prevent potential difficulty in interpreting results due to different numerosity, plots are used where the y-axis is normalised. In this manner, discrepancies in size are smoothed out.

1. **Gender Distribution**

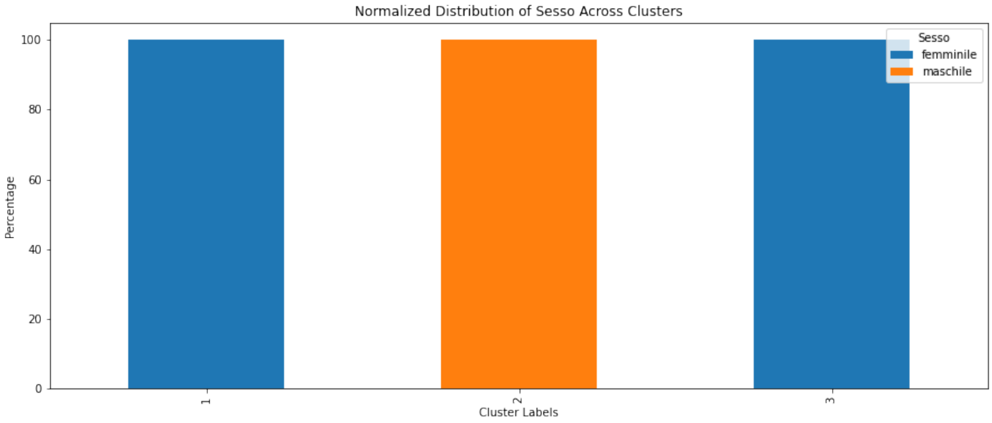


Figure 6.2: Gender distribution across clusters.

Figure 6.2 indicates the gender distribution across the clusters. The male and female genders are sharply divided. Clusters 1 and 3 are made up entirely of women, while cluster 2 is made up entirely of men.

2. **User Type**

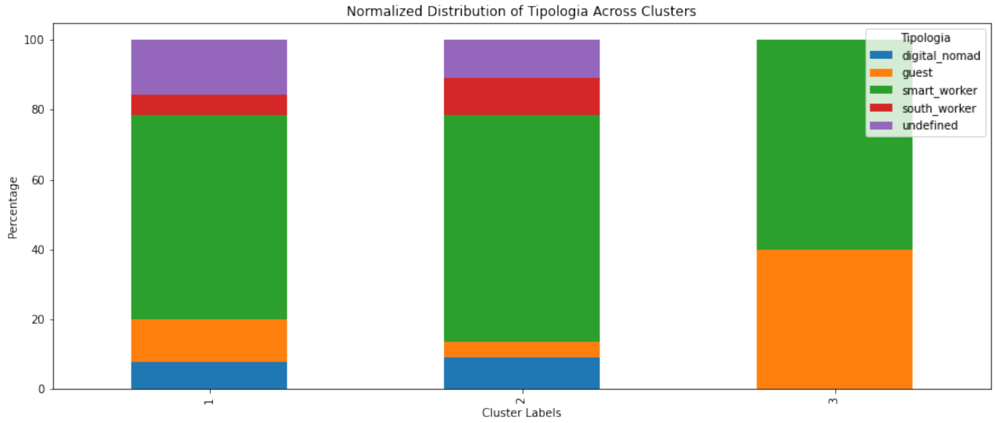


Figure 6.3: User type distribution across clusters.

In Cluster 1, there's a relatively even distribution among four user types: *digital nomad*, guest, *smart worker*, and *south worker*. The undefined category appears to be not so relevant. This suggests that Cluster 1 is quite diverse with a good representation across different user types. The most represented user type is: *smart worker*

Cluster 2 shows a dominance of *smart worker*, taking up more than half of the cluster's composition, followed by a significant portion of guest, and smaller segments of *digital nomad* and *south worker*. The undefined category is again very small or absent. The prominence of *smart worker* indicates that this cluster may represent users with a professional focus on coworking space.

Cluster 3 is markedly different from the other two, with a vast majority being *smart worker*, and only a small portion represented by *digital nomad*. The overwhelming majority of *smart worker* in Cluster 3 suggests that this cluster could be associated with users who frequently use the coworking space for professional purposes. Interestingly, a high percentage of guests is registered.

3. **Work Sector**

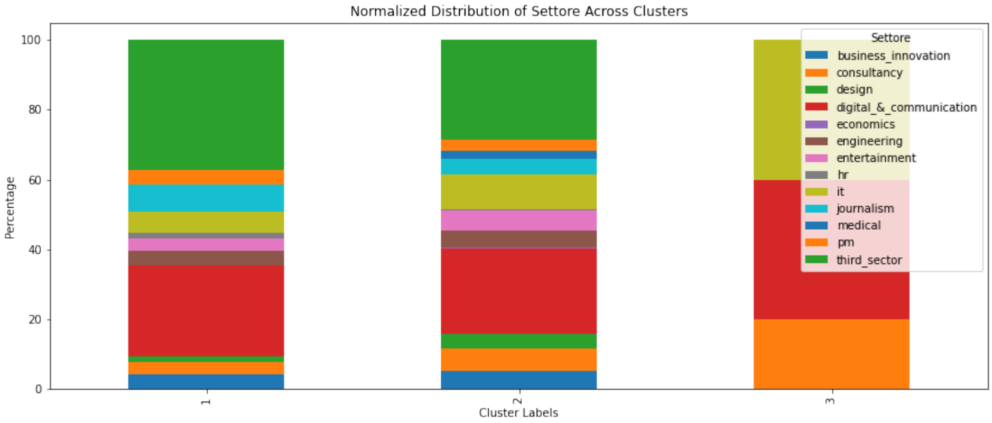


Figure 6.4: Work sector distribution across clusters.

In Cluster 1, there's a diverse mix of sectors, where *digital and communication* and *third sector* seem to be the most dominant ones. This suggests a well-rounded variety of users from different professional backgrounds within this cluster.

Cluster 2 has a different composition, where one sector, potentially the *digital and communication* sector, appears to be significantly more prevalent than others. This may indicate a cluster with a specialized focus or a common professional interest among its users.

Cluster 3 shows a stark contrast, with one work sector, *digital and communication*, dominating the cluster. This indicates a highly specialized user group whose professional activities are concentrated in a single sector.

4. **Origin**

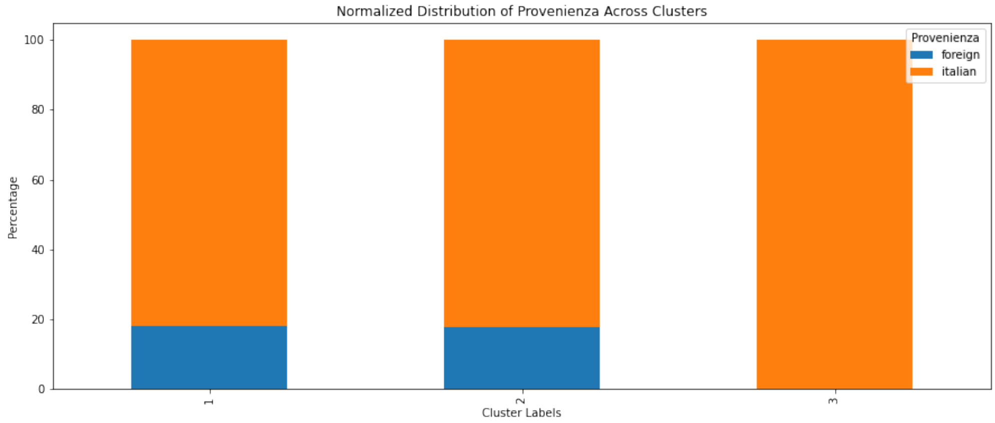


Figure 6.5: User origin distribution across clusters.

Figure 6.5 shows the normalized distribution of nationality across the three clusters. In all three clusters, the majority of users are Italian, as indicated by the larger orange segment. The foreign users, represented by the blue segment at the bottom of each bar, constitute a smaller proportion in each cluster. This suggests that the user base across all clusters predominantly consists of Italian nationals, with international users making up a smaller part of the community.

5. **Access Frequency**

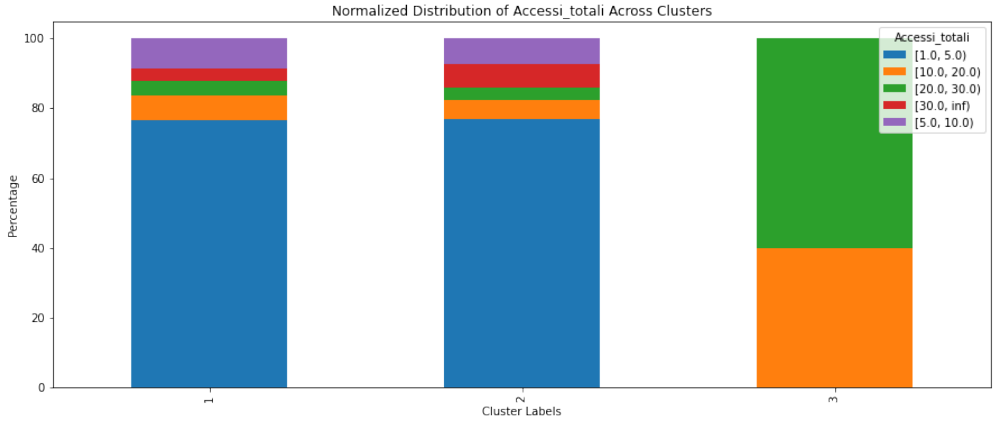


Figure 6.6: Access frequency distribution across clusters.

In Cluster 1, there is a varied mix of access frequencies, although [1.0, 5.0) dominates. This indicates a cluster with users who have low frequency in their use of the coworking space.

Cluster 2 shows a similar composition of Cluster 1 where one particular range of access frequency, [1.0, 5.0) range, is notably more prevalent than the others. This indicates a cluster with users who have low to moderate frequency in their use of the coworking space.

Cluster 3 is characterized by a substantial portion of users falling into the highest frequency range, as indicated by the significant green segment, possibly representing the [20.0, 30.0) range. This suggests that this cluster mainly consists of the most frequent users of the coworking space.

6. **Average Payment**

   The boxplots in Fig 6.7 display the payment distributions.

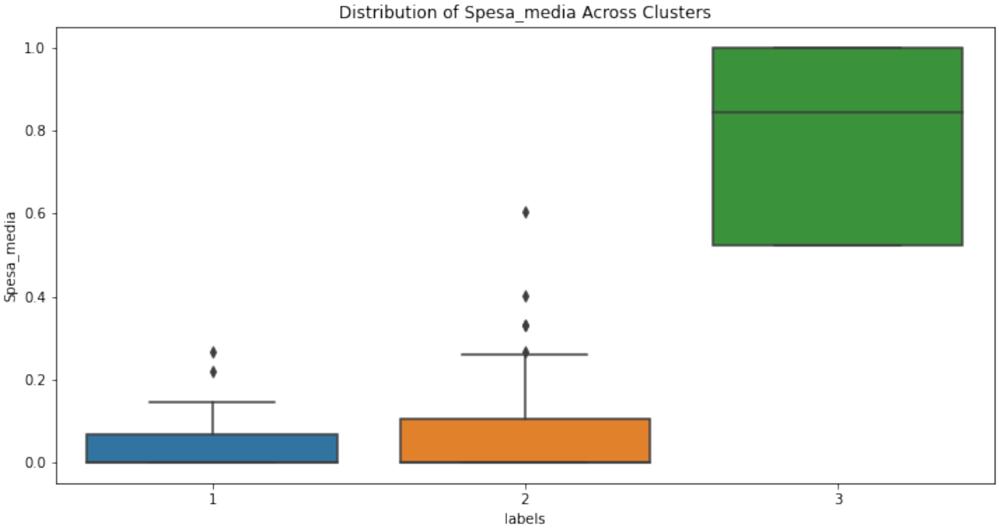


Figure 6.7: Average payment distribution across clusters.

In Cluster 1, represented by the blue boxplot, the distribution is tightly packed around a lower median, indicating that users in this cluster tend to make lower average payments. The presence of a few outliers suggests that there are exceptions within the cluster where some users make significantly higher payments.

Cluster 2, represented by the orange boxplot, shows a slightly higher median average payment compared to Cluster 1, with a broader interquartile range, indicating greater variability in payment amounts among users. The cluster also includes outliers on the higher end, similar to Cluster 1.

Cluster 3, depicted with a green boxplot, stands out with a substantially higher median payment, indicating that users in this cluster tend to spend more on average. The interquartile range is relatively compact, suggesting consistent spending behavior among the cluster's users. There are no apparent outliers in this cluster, which points to a more homogeneous spending pattern.

7. **Payment Variability**

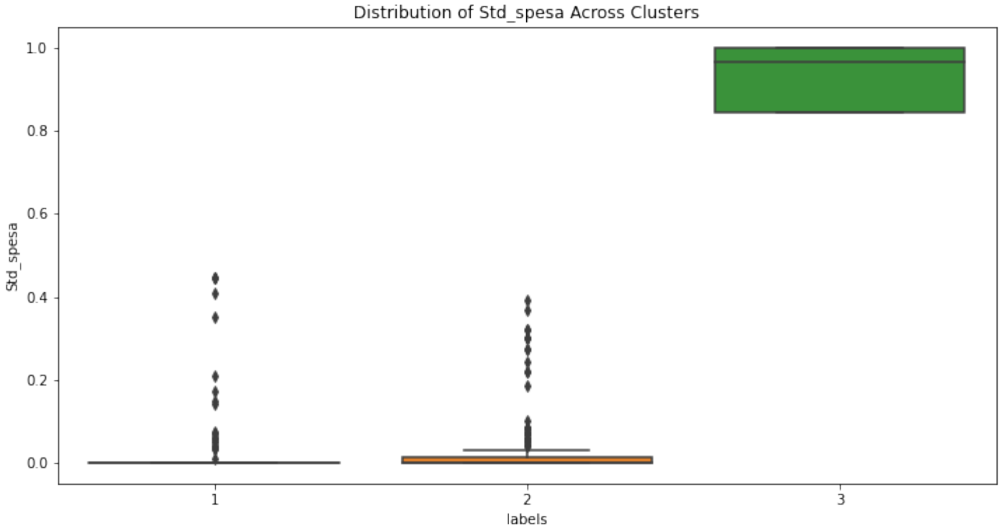


Figure 6.8: Variability distribution across clusters.

Cluster 1 shows a very small median variability in payments, with a compact interquartile range, indicating consistent spending behavior among users. However, there are numerous outliers, suggesting that while most users in this cluster have similar spending patterns, there are exceptions with more variable spending.

Cluster 2 has a slightly larger median variability in payments than Cluster 1, but still indicates relatively consistent spending behavior. There are fewer outliers compared to Cluster 1, which implies more homogeneity in spending patterns among users in this cluster.

Cluster 3 is characterized by a much larger median payment variability, with a broader interquartile range, indicating a diverse range of spending behaviors among its users. The absence of outliers suggests that the variability is integral to this cluster's behavior rather than being driven by a few atypical users.

6.2 Future Directions for Isola Catania's Data Strategy

- **Elevating Temporal Data Insights**

  By refining its approach to temporal analysis, Isola Catania aims to improve its data strategy. The goal is to collect detailed usage data that categorises user activities by exact hours and outlines frequency patterns throughout the week. Adopting advanced sensor-based technologies or digital check-in systems can produce high-quality data, delivering insights into the subtle ways in which different time periods affect user behaviours and overall space use. This enhanced data precision will allow Isola to refine its opening hours, streamline resource allocation, and forecast occupancy trends for future planning. Capturing user activity at specific time slots and across different days enables a thorough examination of peak and quiet periods. These valuable insights are vital for Isola to effectively align its space and resources with the genuine rhythms of user activity.

- **Enriching User Profiles**

  The development of a comprehensive profiling system is crucial for an in-depth comprehension of user interaction across Isola's diverse range of services. The aim is to amalgamate data to create all-encompassing user profiles that depict not only the frequency and duration of coworking space usage but also the users' involvement in supplementary activities and initiatives. Centralising such data enables Isola to identify and interpret user behaviour patterns, thus facilitating strategic decision-making about event attraction and service customisation that resonate with users' distinct preferences and professional pursuits. This approach will also enhance targeted marketing strategies and tailor services to meet specific user needs.

- **Optimizing Event Coordination with Data**

  There are many opportunities for Isola to increase the interconnection between these two areas by analysing the intersection of coworking and event attendance. A detailed examination of event attendance and coworking trends can significantly enhance Isola's event planning. By identifying the correlation between these two factors, Isola can plan its events in a way that captures maximum attendance, particularly if the data reveals that certain user groups tend to attend evening events more frequently.

- **Embracing Predictive Analytics**

The incorporation of predictive analytics into Isola's operations will transform its capability to anticipate and respond to future service needs and event participation. Analyzing historical data with sophisticated machine learning algorithms will enable Isola not only to foresee an increase in demand for specific services but also to predict attendance rates for upcoming events. These prophetic insights are indispensable for the ideal deployment of staff, provision of exclusive offers, and proactively engaging with users.

- **Investing in State-of-the-Art Data Infrastructure**

  Addressing the complexity and volume of data requires a substantial investment in dependable data infrastructure. Isola can manage large-scale data, undertake intricate analyses, and incorporate AI and machine learning innovations by upgrading to advanced analytical tools and leveraging cloud-based platforms. This enhancement will empower Isola to process data in real-time and promptly take action on insights, thereby enhancing operational effectiveness and improving the user experience.

- **Building Analytical Competence**

  Isola's dedication to equipping its team with strong data analytics capabilities is crucial. This involves not only organizing periodic training sessions or workshops but also potentially collaborating with academic institutions. By developing a skilled team that can interpret data and identify data collection opportunities, Isola can take advantage of every user interaction as an opportunity for learning and improvement.

In summary, the proposed enhancements to the data management practices at Isola are carefully designed to create a robust, data-driven framework that supports agile decision-making and enriches the user experience.

Conclusions

Reflecting on the project's results and findings, it is evident that integrating diverse data sources into a unified framework was a significant achievement for Isola. This integration facilitated a comprehensive understanding of Isola's operational environment and user dynamics. The scalable nature of the Data Lake ensures that Isola is well-positioned to handle future data expansions and complexities.

The cluster analysis, employing Hierarchical and K-prototypes clustering methods, was another important component of this project. The study provided a deeper understanding of Isola's user base, allowing for effective segmentation. The identification of user segments such as *Modest Spenders*, *Minimalist Users*, and *Frequent and High Spenders* provided invaluable data for tailoring services, marketing strategies, and enhancing overall user satisfaction. Among the most significant results of this analysis was an in-depth understanding of user behaviour within the co-working area. The research highlighted distinctive patterns of interaction with the space between different demographics, their period of peak utilisation, and their preferences for different amenities. The acquired insights are vital in enhancing the coworking space experience and streamlining its management to meet the changing requirements of its users.

Looking forward, there are several areas where the Data Lake and analytics processes at Isola could be enhanced. Incorporating additional data sources, such as real-time user feedback or more detailed time usage data, could provide even deeper insights into user behavior and preferences. The continuous refinement of clustering algorithms and exploration of new data analysis techniques would undoubtedly enhance the accuracy and relevance of the findings. Furthermore, developing a more interactive and user-friendly interface for data querying and visualization would make the insights more accessible, extending their utility beyond technical teams to a broader range of stakeholders within Isola. In conclusion, this thesis project has demonstrated the immense value and potential of a well-constructed Data Lake and advanced data analysis methods in driving informed decision-making and enhancing user experiences. The insights gleaned from this project have the potential to significantly contribute to the ongoing success and growth of Isola, highlighting the indispensable role of data in shaping the future of business and service delivery.

List of Abbreviations

Acronym	**Definition**
AWS	Amazon Web Services
DDO	Data-Driven Organisation
D^3M	Data-Driven Decision-Making
ETL	Extract, Transform, and Load
HAC	Hierarchical Agglomerative Clustering
HDFS	Hadoop Distributed File System
S3	Amazon Simple Storage Service
SMEs	Small-Medium Enterprises

Bibliography

[1] Aldenderfer, M. S. Cluster Analysis. *Quantitative Applications in the Social Sciences, Sage University Papers, ISSN 0149-192X, Edizione 44.*

[2] Aldenderfer, M. S., & Blashfield, R. K. (1984). Cluster Analysis. *Sage Publications.*

[3] Davenport, T. H., & Bean, R. (2018). Big data and AI strategies: machine learning and alternative data approach to investing. *Journal of Financial Data Science.*

[4] Djokic, N., Salai, S., Kovac-Znidersic, R., Djokic, I., & Tomic, G. (2013). The Use of Conjoint and Cluster Analysis For Preference-Based Market Segmentation. *Inzinerine Ekonomika-Engineering Economics*, 24(4), 343-355.

[5] Ezugwu, A. E., Ikotun, A. M., Oyelade, O. O., Abualigah, L., Agushaka, J. O., Eke, C. I., & Akinyelu, A. A. (2022). A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects. *Engineering Applications of Artificial Intelligence, 110*, 104743.

[6] Fischer, H., Wiener, M., Strahringer, S., Kotlarsky, J., & Bley, K. (2022). From Knowing to Data-Driven Organizations: Review and Conceptual Framework. *Australasian Conference on Information Systems.*

[7] Ghosal, A., Nandy, A., Das, A. K., Goswami, S., & Panday, M. (2019). A short review on different clustering techniques and their applications. In *Advances in Intelligent Systems and Computing* (pp. 69–83).

[8] Hartmann, P. M., Zaki, M., Feldmann, N., & Neely, A. (2016). Capturing value from big data–a taxonomy of data-driven business models used by start-up firms. *International Journal of Operations & Production Management.*

[9] Huang, Z. (1998). Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values. *Data Mining and Knowledge Discovery, 2*, 283–304. https://doi.org/10.1023/A:1009769707641

[10] Hukkeri, T. S., Kanoria, V., & Shetty, J. (2020). *A Study of Enterprise Data Lake Solutions.* International Research Journal of Engineering and Technology (IRJET), 07(05), p.1.

[11] Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering. *ACM Computing Surveys, 31*(3), 264–323.

[12] Johnson, S.C. (1967). Hierarchical Clustering Schemes. *Psychometrika, 32*(3), 241–254.

[13] Kardes, F. R. (1999). Consumer Behavior and Managerial Decision Making. *Addison-Wesley.*

[14] Kardes, F. R. (1998). Consumer behavior and managerial decision making. http://ci.nii.ac.jp/ncid/BA56272905

[15] L'Esteve, R.C. (2023). Designing a Secure Data Lake. In: *The Cloud Leader's Handbook.* Apress, Berkeley, CA. https://doi.org/10.1007/978-1-4842-9526-7_11

[16] McAfee, A., & Brynjolfsson, E. (2012). Big data: the management revolution. *Harvard Business Review*, 90(10), 60-68.

[17] Miloslavskaya, N., & Tolstoy, A. (2016). Big Data, Fast Data and Data Lake Concepts. *Procedia Computer Science*, 88, 300-305.

[18] Nambiar, A., & Mundra, D. (2022). An Overview of Data Warehouse and Data Lake in Modern Enterprise Data Management. *Big Data Cogn. Comput.*, 6, 132.

[19] Punj, G., & Stewart, D. W. (1983). Cluster Analysis in Marketing Research: Review and Suggestions for Application. *Journal of Marketing Research, 20*(2), 134. https://doi.org/10.2307/3151680

[20] Řezanková, Hana, Praze, Vysoká, & Praha,. (2009). Cluster analysis and categorical data. *Statistika, 89.*

[21] Runiewicz-Wardyn, M. (2014). Geographic and technological pattern of knowledge spillovers as evidenced by technical universities in CEE countries. *The Engineering Economics, 25*(4). https://doi.org/10.5755/j01.ee.25.4.3758

[22] Xu, D., & Tian, Y. (2015). A comprehensive survey of clustering algorithms. *Annals of Data Science, 2*(2), 165–193.

[23] Khine, Pwint & Wang, Zhao. (2018). *Data lake: a new ideology in big data era.* ITM Web of Conferences, 17, 03025.

List of websites

[24] *Isola Catania on Facebook* www.facebook.com/Isola.Catania

[25] *Isola Catania on Instagram* www.instagram.com/isola.catania/

[26] *Isola Catania* www.isola.catania.it

[27] *Isola Catania on YouTube* www.youtube.com/@isolacatania

[28] *Isola Catania on LinkedIn* www.linkedin.com/company/isola-catania/

[29] *Visual Studio Code* https://code.visualstudio.com/

[30] *Kmodes on GitHub* https://github.com/nicodv/kmodes

[31] *SciPy on GitHub* https://github.com/scipy/scipy

[32] *Python* https://www.python.org/