A picture containing logo

Description automatically generated

All work must be your own work and any sources used must be referenced!

NCFE Level 3 Certificate in Data

Unit 01 Understand how to source data

|  |  |
| --- | --- |
| **Name:** | OUISSAL MESKINE |
| **Trainer:** | Samuel Mayes |
| **Course ID:** | 603/7882/7 |

**Table of Contents**

[Learning Outcome 1: Understand how to source data 1](#_Toc164092953)

[1.1 Explain the role of data in the context of a digital world 1](#_Toc164092954)

[1.2 Explain the internal datasets that can be found for analysis 1](#_Toc164092955)

[1.3 Explain the external datasets that can be found for analysis 1](#_Toc164092956)

[1.4 Describe the open datasets that can be found for analysis 2](#_Toc164092957)

[Learning Outcome 2: Understand data formats and their importance for analysis for a specific business requirement. 3](#_Toc164092958)

[2.1 Explain the functions and purpose of data formats 3](#_Toc164092959)

[2.2 Selecting the appropriate data format 3](#_Toc164092960)

[Learning Outcome 3: Purpose, principles and functions of data architecture for a specific business requirement 4](#_Toc164092961)

[3.1 Describe the purpose of data architecture: 4](#_Toc164092962)

[3.2 Purpose and function of integration 5](#_Toc164092963)

[3.3 Access Requirements 5](#_Toc164092964)

## Learning Outcome 1: Understand how to source data

|  |
| --- |
| 1.1 Explain the role of data in the context of a digital world |
| In your write up include:  How data underpins every digital interaction and connection across the digital landscape:  Data represent phenomena, events, or facts using symbols, expressions, or media not yet interpreted or used for a specific function.  When we input data for analysis, the goal is to derive information by extracting meaning and insight from a single piece of data or, more commonly, from a set of data, or *BIG DATA*, a very large amount of data*.*  The data will be subjected to an interpretation process during which this information will become relevant, valued and knowledgeable, comprehensible to the reader or the recipient who receives it.  In brief, this can be placed into context as:     1. DATA > INFORMATION > KNOWLEDGE   In our fast-paced daily lives, we are quickly becoming more integrated into the digital world, which is becoming an essential part of how we connect and progress in today's technology-driven society. The large volume of data produced daily defines what we now refer to as the digital modern lifestyle. This digitalization phenomenon not only produces but also encompasses the process of converting diverse types of data into digitalized forms, utilizing a myriad of tools, media, software, and hardware, which are becoming numerous every day, such as Big Data.  *Big data* refers to an enormous amount of data that cannot be analyzed using traditional methods. It includes :   * structured (like databases) * Unstructured data (such as social media posts, videos, and images).   *Big Data* is also defined by other factors:   * Velocity: Refers to the frequency and speed with which data is generated and collected. * Variety : Refers to the different types of data that are collected. * Value : Extraction of meaningful information from data * Veracity: Generation of data in real time or close to real time ( transaction) * Volume : large amount of data   Big data offers enormous potential as it allows us to uncover patterns and trends that were previously unseen. This can lead to new discoveries and insights. For instance, companies can utilize this vast amount of data to predict consumer trends and customize their products, while scientists can use big data for large-scale research to uncover new information about our world.  This can also be placed into context as:  2. DATA > INFORMATION > KNOWLEDGE> WISDOM  Certainly, a substantial volume of data can also offer numerous advantages to businesses. With big data, companies can swiftly engage with their consumers by capturing their online behavior to comprehend their preferences, needs, and habits. Subsequently, they can tailor personalized strategies based on this information. Data analysts play a pivotal role in this system by not only *saving* and organizing the data but also by *leveraging* it to drive business growth and *profitability*:    3. SAVE MONEY -> PROTECT MONEY –> MAKE MONEY  There are numerous examples of how big data is used today. Here are some of the most common :   * Social media * Marketing * Logistics * Healthcare * Business * Finance * Research and development * Smart cities * Logistic * Pick an example which goes through a user’s journey where transactional and booking data is stored, processed and how the information is then used. Make sure you include the points below: * Customer centricity * Transactional data, including purchasing, think about invoices, statements, etc. * Booking data, including reservations, think about availability and peak/off-peak pricing.   Example: *A User schedules a spa session using an online platform or a spa app.*  User can explore various spa treatments and massage options by visiting the website or using the app. To make the user experience even better, the platform uses cookies and his past search history to provide personalized suggestions. For instance, if he has booked relaxing massages before, the homepage may display exclusive deals specifically for massages based on his preferences.  =>The concept ofcustomer centricity is a strategic approach, prioritizing the customer's needs and preferences above all else. This involves creating a seamless and enjoyable customer experience and building lasting relationships with the customer base. In this strategy, the focus shifts to the customer as the center of interest rather than the service provided.  in this example data will be collected as :   * Demographic data: Collect personal data of customers (age, gender, Email, etc.) via online or on-site registration. * Behavioral Data: Track customer interactions with the website and app, including researches made with specific goals the client is aiming at, pages visited, treatments booked previously. * Customer Feedback: Use post-treatment surveys and reviews to gather opinions and suggestions.   =>When you schedule a massage session at a spa, a variety of transactional data is created and managed from a technical standpoint. This information is essential for operational efficiency, customizing services, and ensuring clarification and confirmation. Some of the transactional data involved are :  Purchasing details : user schedules a massage session, and the platform handles payment, creates an invoice, and stores transaction data such as:   * Amount paid * Payment method (credit card, debit card, cash, online payment service) * Date and time of payment * Billing address (if applicable) * Confirmation number or booking reference * Notifications sent to the customer (confirmation email, reminder SMS, etc.)   =>When scheduling a spa treatment session, it's important to provide comprehensive details to ensure a personalized and enjoyable experience. This includes specific Therapist Information such as the name of the assigned therapist and their unique ID or employee number, It's important to include the following :  Booking Details:   * Date and time of the appointment. * Type of massage or treatment booked (e.g., Swedish, deep tissue, hot stone) * Duration of the treatment * Special requests or preferences (e.g., specific therapist, room temperature)   =>As well as any promotional codes or discounts can be added to optimize customer satisfaction.  If a User is a member of a loyalty program, it is made sure to display the relevant details as well in an invoice.  Additionally, specification of any extra services a user would like to include, such as aromatherapy or extended session time, as well as any products They wish to purchase, such as skin care items or gift cards, oil treatments...  => When a user books a spa massage session, the platform goes through a series of important processes to ensure a smooth and user-friendly experience.  Firstly, it checks and updates the availability of spa sessions in real-time. It uses mechanisms like database transactions to prevent double bookings and maintain accurate scheduling. As part of this process, the platform temporarily locks the selected slot to ensure its availability until the booking is confirmed.  At the same time, the platform dynamically calculates the price of the session based on predefined peak and off-peak pricing rules, considering factors such as the time of day, weekdays versus weekends, and holidays. This dynamic pricing is displayed to the user, ensuring transparency and allowing for informed decisions. These steps, supported by robust backend technologies and real-time data handling, create a seamless booking experience that efficiently balances user needs with operational integrity.  For each of these, find an example of how the data is collected, processed, and used to gain insight:   * Online applications (for example, access, targeted marketing across applications) * Physical world (for example, location, transactions across multiple sites) * Smart devices (for example, virtual assistant, home management) * Internet of Things (IoT) * Technologies (for example, building management, transportation, manufacturing) * Customer interactions/centricity   A spa massage centre utilizes data warehousing and data mining to gather information about customer preferences, appointment history, and product usage. Customer Relationship Management (CRM) is used to collect all customer data, including information and feedback.  This data is then analyzed to identify patterns and trends, which helps improve services, streamline operations, and personalize customer experiences. For example, providing an :  Online Applications allow users to share their spa experiences directly or through Social Media, increasing the visibility of the spa and encouraging positive word of mouth. Analyzing customer feedback can help in the development of new massage techniques or products.  The Spa centres can use this data to optimize staffing schedules and manage inventory, developing a greater experience inside the Spa.  IoT devices can assist a massage centre in collecting real-time data on room occupancy and equipment usage, as well as in adjusting the environment such as *Smart Thermostats*, *systems equipped with sensors* ensuring a soothing environment managing stocks.  To ensure security, the staff personnel use secure access to rooms for client’s safety.  By collecting customer data with *tablets* to tailor the experience and creating customized products to meet their needs, customer satisfaction levels are increased, ultimately contributing to the achievement of business objectives.  Utilizing technological manufacturing can help a spa advance by offering innovative equipment and products such as heat therapy beds and chairs, We use modern technology to make high-quality skincare products with advanced ingredients that target specific skin concerns, Scanners are also important for personalized treatments. |

|  |
| --- |
| 1.2 Explain the internal datasets that can be found for analysis |
| What are Internal datasets What do they include? Give real life examples, including:   * inventory (for example, sales) * financial * marketing * customer database * HR/personnel   =>An internal Dataset is an organized electronic data store, managed on computers through dedicated software. It uses data modeling techniques to create a well-organized set of data stores. This enables different users to access and work with the data using various programs.  An inventory can provide you with information about the availability of products in a Warehouse:   |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | | **Product Number** | **Product Name** | **Stock** | **price** | **Added item Date** | **Expiration date** | | 0001 | Magnesium | 456 | 7 £ | 01/01/2024 | 01/11/2028 | | 0002 | Calcium | 277 | 5 £ | 03/01/2024 | 03/21/2027 | | 0003 | Vitamin C | 267 | 4 £ | 06/01/2024 | 02/31/2025 | | 0004 | Zinc | 086 | 9 £ | 08/01/2024 | 01/04/2025 |   =>A financial dataset contains various financial metrics and company transactions. This data is crucial for financial planning, analysis, and reporting.  For example, this dataset enables the monitoring of the wellness center's primary financial transactions. It provides a clear overview of the income generated from treatments and product sales, as well as the expenses for rent, product purchases, salaries, and maintenance :   |  |  |  |  | | --- | --- | --- | --- | | **DATE** | **DESCRIPTION TITLE** | **TYPOLOGY** | **AMOUNT** | | 05/01/2024 | Rental | expense | 1000 £ | | 06/01/2024 | Equipment purchased | expense | 1100 £ | | 07/01/2024 | Product Sales | income | 1500 £ | | 07/01/2024 | Product Purchased | expense | 500 £ |     =>This dataset enables the analysis of the effectiveness of various marketing campaigns and channels. By comparing the budget, inquiries generated, and new potential customers, the spa centre can identify which campaigns and channels provide the best return on investment and adjust future marketing strategies.  Here is a simple marketing dataset example for a spa center:   |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | | **Date** | **Theme**  **campaign** | **Marketing Channel** | **Investment** | **Number of inquiries** | **Potential Clients** | |  | New Year pack | Health and Wellness Podcasts | 700 £ | 153 | 80 | | 14/02/2024 | Valentine’s Day | Referral program | 400 £ | 86 | 36 | | 10/03/2024 | Mother’s Day event | Tours and social media | 600 £ | 71 | 27 |   =>This customer dataset serves as a valuable tool for managing customer information effectively, enabling personalized communication, targeted promotions, and comprehensive tracking of customer visits and preferences:   |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | | **Customer** | **Full name** | **Email** | **DOB** | **Phone number** | **Joining date** | **Treatment**  **chosen** | **Last visit** | | 001 | Frank Goldman | XX@gmail.com | X | XXX | XX | XX | XX | | 002 | Laura Pearce | XX@gmail.it | X | XXX | XX | XX | XX | | 003 | Sara Wilson | XX@hotmail.com | X | XXX | XX | XX | XX | | 004 | Greg Ronald | XX@gmail.fr | X | XXX | XX | XX | XX |   =>This HR dataset contains important details about each employee, including their ID, name, position and department, hire date, and salary, specifically customized for a spa environment.   |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | | **Employee ID** | **Full Name** | **Date of hiring** | **department** | **position** | **Salary** | | 002 | Diana Hoover | 02/05/2021 | Wellness | Massage Therapist | 3500 | | 005 | Linda Crook | 03/04/2020 | Management | Spa Manager | 5700 | | 004 | Isabel Ferry | 06/09/2023 | Front Desk | Receptionist | 2650 | | 003 | Riccardo Remo | 01/03/2022 | Wellness | Hairdresser | 3560 | |

|  |
| --- |
| 1.3 Explain the external datasets that can be found for analysis |
| What are External datasets What do they include? Give real life examples, including:   * External organisations * Sharing agreement * Trusted sources   =>An external dataset is a collection of data that comes from outside your own system or organization. It is not created by a single source but by multiple other sources, like government websites, research organizations, or other companies. This Data is used to enhance or compare, some of the examples are :   * Publicly Available Databases * Public API * IoT Networks * Non Profit Organisations * Commercial Data providers   External datasets can be used to gain new insights or add extra information that you did not have before. They help in making better decisions in businesses by providing a broader perspective and make money.  =>These External datasets must consider quality relevance licensing agreements and trustworthiness in the context of privacy and compliance issues following legal and ethical standards   * Example of an External organisation data supplier is The United States Census Bureau provides detailed demographic, economic, and social data for use in business, research, and policymaking. * Example of a Sharing agreement is Apple and Nike collaboration to integrate user activity data from Nike's fitness apps with Apple Health. This allows users to synchronize their fitness data across their iPhone and Apple Watch for a more comprehensive view of their health and fitness activities. * Example of a trusted source: Aeroporti Di Roma in Compliance with GDPR, the (ADR) manages and develops Roma Fiumicino (Leonardo da Vinci) and Ciampino airports. Several airlines operate from these airports, transporting passengers to destinations around the globe. The Airport Data Recorders (ADR) comprehend the importance of reliable data for swiftly grasping and predicting customer behaviours. They also recognise their duty to safeguard their customers' personal data. |

|  |
| --- |
| 1.4 Describe the open datasets that can be found for analysis |
| What are Open datasets Where can you obtain these from: Give real life examples, including:   * Department for Environment, Food and Rural Affairs (DEFRA) * Local government * Office for National Statistics (ONS)     Open data means information collected, produced or paid for by public bodies and made available free of charge for other purposes.  Open data has many benefits. It can help public administrations work better, support businesses to grow, and improve overall social welfare.   * Department for Environment, Food and Rural Affairs (DEFRA) example: Woodland Tree Health grants 2023: Countryside Stewardship   <https://www.gov.uk/government/publications/woodland-tree-health-grants-2023-countryside-stewardship>     * Local government : Birmingham City Council provides access to a variety of datasets related to the city and its operations, including information on demographics, housing, transportation, education, and more.   Here is an example of a pdf     * Office for National Statistics (ONS) :   [Our changing population is there for all to see  | National Statistical (ons.gov.uk)](https://blog.ons.gov.uk/2024/04/12/our-changing-population-is-there-for-all-to-see/)  Add in your breakout room task here! |

## Learning Outcome 2: Understand data formats and their importance for analysis for a specific business requirement.

|  |
| --- |
| 2.1 Explain the functions and purpose of data formats |
| Include:   * Numeric integer * Temporal * Text * Geospatial * Media * Logical * References   Data formats are like rules for organizing and sharing information between different systems. They ensure that data is understood correctly by defining its structure and representation. Following these formats helps businesses ensure that their technology can work together smoothly, making information exchange easier and more efficient :  -Numeric integer formats: handle whole numbers for precise calculations.  -Temporal formats: manage dates and times for scheduling and event tracking.  -Text formats: structure textual data for efficient storage and retrieval.  -Geospatial formats: encode location data for mapping and geographic analysis.    -Media formats: handle multimedia content like images and videos.  -Logical formats: manage Boolean values for decision-making.  -Reference formats: link data to external sources, facilitating integration and validation across systems.  =>Each format serves a crucial purpose in ensuring data is structured and utilized appropriately.  Add in your Spreadsheets here as evidence of what data types you have used and why they are appropriate. |

|  |
| --- |
| 2.2 Selecting the appropriate data format |
| Include:   * Consistency (using the same data format when merging datasets, knowing what data you want) * Calculation (using the appropriate format, knowing what you will do with the data) * Conversion (converting data from one type to another for analysis; for example, costs stored as text rather than a number)   Add in your Spreadsheets here as evidence for all the formula and formatting you have used and why they are appropriate.   * Sums - BODMAS * IF Statement * Conditional Formatting * VLOOKUP * AVERAGE * COUNTIF * MIN/MAX   *The spreadsheet have all the formula above stated in form of exercises based on the workbooks we did during the course*        Use the lesson PowerPoint to help you to ensure all content is covered. |

## Learning Outcome 3: Purpose, principles and functions of data architecture for a specific business requirement

|  |
| --- |
| 3.1 Describe the purpose of data architecture: |
| Include and give examples for:  • The framework guiding the development and operation of information systems and data storage  • The rules and policies that can define and explain the type of data  Data Architecture is intended to provide organise and design consistent framework for managing, storing, processing and accessing across an organisation. It aims to ensure that data are accurate secure and ready to use efficiently.   * The BRD, ( Business Requirement Document) is a guideline project management tool in which it is determined the HOW, WHERE, WHO, WHAT,WHY, WHEN,by a business analyst to explain the undergoing of a project or a business requirement. * Testing plays a crucial role in the integration and the functionality to test the results that are expected * Evaluate the factors that impact the data architecture based on access and security ( CIA Triad ) which ensures the right measures are set up for data protection : Confidentiality, Integrity and Availability   For example, in a retail company such as Amazon or Costa Coffee In a retail company, data architecture determines how sales data from various stores is gathered, stored in databases, and accessed for analysis. This ensures efficient organization of data, facilitating easy retrieval, trend analysis, and informed decision-making. |

|  |
| --- |
| Describe the principles of data architecture: |
| Include and give examples for:   * Access – data available for user functions * Definition – data is valued as an asset * Managed – data is in a form which facilitates maintenance and understanding of the   data pipeline process   * Secured – data has the appropriate security controls applied and is only accessed by   appropriate users   * Shared – data can be extracted and shared between communities, without   compromising safety or exporting sensitive information  The Principle of data architecture is to have easy access to data but ensure to protect these data from unauthorized access as well as define the typology and the relationship attributed to the data to ensure it is used correctly and understood.  Maintaining data ensures it is always up to date and improved to remain accurate and valuable a methodology used would be ( PESTLE) analysis.  Security is to secure data and protect it from unauthorized access and threats etc., another principle is when it comes to share data we should analyst consider ETLVisualization, by underpinning data to enable collaboration and make it available for it to be understood in different ways for different purposes .  For example, in a Bank, the data architecture framework might include a centralized data warehouse for customer transactions from various platforms and methods used by customers,: Access and Management.  Financial advisors use customer data to recommend personalized investment options: Definition.  Secured and: Data governance policies to ensure legal guidelines are followed, and methods to secure sensitive financial data such as converting information into a secret code to protect it, encryption, Only someone with the right key or password can decode (decrypt) it and read the original information. This framework supports risk management, fraud detection, and personalized customer services. Shared: Aggregate financial market trends are shared with investors through secure reports without revealing individual client details. |

|  |
| --- |
| Describe the functions of data architecture: |
| Include and give examples for:   * Organise data – grouped by selected criteria:   + Data types (for example, string (or str or text), character (or char), integer (or int),   + Float (or real), boolean (or bool))   + Formats (for example, database, spreadsheets, comma-separated values (CSV) file) * Data storage – specifying the different types of data storage and its location:   + On premises   + Cloud   + Third-party   + Hybrid * Permissions and access across different systems (for example, file server):   + Levels of permissions   + Levels of access   + Multi-factor authentication * Data structures:   + Structured   + Unstructured   + Semi-structured  Data architecture arranges how data is stored and used, making sure it's organized and secure for easy access and accurate analysis by implementing and designing data system for this. It helps businesses manage information effectively for better decision-making. Organise Data:   * Data types: string: such as a customer name or student   integer (or int): quantity of some products     * Float (or real): price of the products Boolean (or bool): true or false identifying availability or unavailability * Formats: database such as SQL, spreadsheets like Excel   Data storage:   * + On-premises: servers   + Cloud: azure   + Third-party: when a company hires a cloud service provider   + Hybrid: a company that uses her own servers and a cloud service   Permissions and access across different systems:   * + Levels of permissions: Admins have full access, while users have limited rights to specific folders.   + Levels of access: In a project management tool, team members have read-write access to tasks to update but clients, on the other hand, are given read-only access to project progress reports, allowing them to view but not edit project details.   + Multi-factor authentication: Requires users to verify identity with a password and a code sent to their phone.   Data structures:   * + Structured: CVS docs   + Unstructured: videos and images jpg   + Semi-structured: websites, social media posts |

|  |
| --- |
| 3.2 Purpose and function of integration |
| Include and give examples for:   * Cost * Size of organisation * Third-party applications * Benefits such as efficiency   Integration refers to the process of combining data from different sources into a unified system or structure:  Cost: to avoid manual data entry into software by automating data transfer between systems for example in an online store.  Size of integration: Multinational companies integrate multiple systems to manage vast amount of data.  Third-party application: when applications are introduced as third parties to upskill functionalities like PayPal to secure easy transactions between two parts.  Benefits such as efficiency: to ensure data updates are available in a real time basis this type of integration improves efficiency in a warehouse for example, to prevent stockouts. |

|  |
| --- |
| 3.3 Access Requirements |
| Include and give examples for:   * User * Systems   User and Systems:  The ACID principle stands for:   * Atomicity: when a user makes a transaction the Data records it in the right order and is reflected directly into the system. * Consistent: it is of extreme importance to maintain rules applications solid consistently to avoid inconsistency during data entry. * Isolate: the user inputs should hold accountability once are into a system or for example when a group of users are working in a shared drive it is important to isolate one transaction from another to avoid merging and blockages. * Durability: Ensure an organised data record of past transactions and check them up when needed for example a student record.   -Both user and system should put safeguards to ensure security as an access requirement  -Data cannot be visualized sometimes through direct access it for users and systems, some need systems like API and FTP.  -Some Systems are mandatory to implement if the number of users increases in this case to avoid the impact of the scalability on the data architecture it may need modification to answer and accommodate the demands  -BRD Business Requirement Document is a vital management tool that ensures systems the access to some data before some other data . |

|  |
| --- |
| Security Requirements |
| Include and give examples for:   * Confidentiality * Integrity * Availability   The Mandalov Matrix integrates with the CIA Triad by aligning business goals with security measures to ensure confidentiality, integrity, and availability. For example, in a healthcare system:   * Confidentiality: Encrypting patient records ensures only authorized medical staff can access them. Avoiding *Disclosure* * Integrity: Implementing data validation checks prevents unauthorized changes to patient information. Avoiding *Alteration* * Availability: Using redundant servers and backups guarantees patient data is accessible even during system failures. Avoiding *Disruption* |