Professional Certificate in Data Quality Assurance using AI in Education

Data Collection and Integration

Artificial Intelligence (AI): the simulation of human intelligence in machines that are programmed to think like humans and mimic their actions. AI can be categorized as either weak or strong. Weak AI, also known as narrow AI, is an AI system that is designed and trained for a particular task. Virtual personal assistants, such as Apple's Siri, are a form of weak AI. Strong AI, also known as artificial general intelligence, is an AI system with generalized human cognitive abilities. When presented with an unfamiliar task, a strong AI system is able to find a solution without human intervention.

Data Collection: the process of gathering and measuring information on variables of interest, in an established systematic fashion that enables one to answer stated research questions, test hypotheses, and evaluate outcomes. Data collection is a vital step in the data quality assurance process, as it lays the foundation for any analysis that will be performed.

Data Cleaning: the process of preparing data for analysis by removing or modifying data that is incorrect, incomplete, irrelevant, duplicated, or improperly formatted. Data cleaning is an important step in the data quality assurance process, as it helps to ensure that the data used for analysis is accurate and reliable.

Data Governance: the overall management of the availability, usability, integrity, and security of data. Effective data governance ensures that data is used in a consistent and controlled manner, and that it is protected from unauthorized access or use.

Data Integration: the process of combining data from different sources into a unified view. Data integration is a key challenge in many industries, as it requires the ability to access, clean, transform, and load data from a variety of sources into a single system.

Data Lake: a large storage repository that holds a vast amount of raw data in its native format until it is needed. Data lakes are designed to handle high volumes of data from a variety of sources, and are often used in big data and AI applications.

Data Mart: a subset of an organization's data that is designed to serve a specific business unit or group of users. Data marts are typically smaller than data warehouses, and are focused on providing fast, easy access to data for a specific purpose.

Data Mining: the process of discovering patterns and knowledge from large amounts of data. Data mining uses a variety of techniques, including machine learning, statistics, and databases, to extract insights from data.

Data Model: a conceptual representation of data structures and the relationships between them. Data models are used to design and implement databases, data warehouses, and other data storage systems.

Data Quality: the degree to which data is accurate, complete, consistent, and timely. Data quality is an

important consideration in any data-driven activity, as poor quality data can lead to incorrect conclusions and poor decision-making.

Data Quality Assurance: the process of ensuring that data is of high quality and fit for its intended use. Data quality assurance includes a range of activities, such as data profiling, data cleaning, data validation, and data monitoring.

Data Science: an interdisciplinary field that uses scientific methods, processes, algorithms, and systems to extract knowledge and insights from structured and unstructured data. Data science is a rapidly growing field, driven by the increasing availability of data and the need for organizations to make data-driven decisions.

Data Warehouse: a large, centralized repository of data that is designed to support business intelligence and decision-making. Data warehouses typically store data from multiple sources, and are used to provide a unified view of an organization's data.

Deep Learning: a subset of machine learning that is inspired by the structure and function of the brain, specifically the interconnecting of many neurons. Deep learning models are able to learn and improve from experience, and are often used in AI applications, such as image and speech recognition.

Extract, Transform, Load (ETL): the process of extracting data from various sources, transforming it into a suitable format, and loading it into a target system, such as a data warehouse. ETL is a key step in the data integration process, as it enables data to be combined from multiple sources into a single system.

Machine Learning: a type of artificial intelligence that allows systems to learn and improve from experience without being explicitly programmed. Machine learning algorithms use statistical methods to identify patterns in data, and can be used to make predictions or decisions based on that data.

Natural Language Processing (NLP): a field of artificial intelligence that focuses on the interaction between computers and human language. NLP enables machines to understand, interpret, and generate human language in a valuable way, and is often used in applications such as virtual personal assistants, chatbots, and language translation.

Predictive Analytics: the use of statistical algorithms and machine learning techniques to identify the likelihood of future outcomes based on historical data. Predictive analytics is used in a wide range of industries, including finance, healthcare, and marketing, to make informed decisions and improve outcomes.

Structured Data: data that is organized in a predefined manner, often in a tabular format, and can be easily searched, sorted, and analyzed. Structured data is typically stored in databases, and includes information such as customer names, addresses, and purchase history.

Unstructured Data: data that does not have a predefined structure, and is often text-heavy, such as emails, social media posts, and documents. Unstructured data is more difficult to search, sort, and analyze than structured data, but can provide valuable insights when properly processed and analyzed.

Web Scraping: the process of automatically extracting information from websites. Web scraping is often used to gather large amounts of data from the web for analysis, and can be performed using a variety of tools and techniques.

This glossary provides an overview of the key terms and concepts related to data collection and integration in the context of the Professional Certificate in Data Quality Assurance using AI in Education. Understanding these terms is essential for anyone working with data, as they provide a common language and framework for discussing and working with data. By mastering these concepts, learners will be well-prepared to tackle the challenges and opportunities presented by the ever-growing amount of data available in today's world.