Data Provenance:

Provenance (from the French provenir, 'to come from/forth') is the chronology of the ownership, custody or location of a historical object. The term was originally mostly used in relation to works of art but is now used in similar senses in a wide range of fields, including archaeology, paleontology, archives, manuscripts, printed books, economy, science and computing. The primary purpose of tracing the provenance of an object or entity is normally to provide contextual and circumstantial evidence for its original production or discovery, by establishing its later history, especially the sequences of its formal ownership, custody and places of storage. Provenance is conceptually comparable to the legal term *chain of custody,* it is also known as *lineage* and is sometimes synonymous to *traceability* in supply chain management, software development, healthcare and security.

However within computer science, people use the term "provenance” to mean the lineage of data, as per *data provenance*, with research in the last decade extending the conceptual model of causality and relation to include processes that act on data and agents that are responsible for those processes. *Data provenance* refers to a method or technique that can be used to record data origin and/or data operation and processing history in various applications. Provenance data/information is a kind of metadata that records data origin and/or data processing history that can be collected by a *data provenance* technique.

With *data provenance*, it can be inferred who is accountable for the modification of data, how and where it happened and which other data influenced the process of creating new pieces of data. In data protection, *data provenance* can be used to enable the data subject to carry out his right to information. In European data protection law, everybody has the right to know where the organisation accountable got his data from, what the data was used for, where it was transferred to and how long it is stored. Only by knowing the exact data flow to and from the organisation accountable, it can be assured that this information can be provided.

Some researchers separate *data provenance* in two main types: *source* provenance and *transformation* provenance. *Source* provenance will tell the user who is responsible for the source of the data, ideally including information on the originator of the data. *Transformation* provenance gives details regarding how the data has been used and modified and often includes information on how to cite the data source or sources. *Data provenance* is of particular concern with electronic data, as data sets are often modified and copied without proper citation or acknowledgement of the originating data set. Databases make it easy to select specific information from data sets and merge this data with other data sources without any documentation of how the data was obtained or how it was modified from the original data set or sets. The automated analysis of *data provenance* has been described as a means to verify compliance with regulations regarding data usage such as those introduced by the EU GDPR.

By having knowledge of records of the inputs, entities, systems, and processes that influence data of interest and providing a historical record of the data and its origins, *data provenance* can enable the support of forensic activities such as data-dependency analysis, error/compromise detection and recovery, auditing, compliance analysis, assess quality and trustworthiness of data, improve data readability or solve copyright issues.

Requirements*:*
A data provenance approach/technology should have **[Requirement]** *that* [Description].

Logging: Provide data flow trace resulting from the transfer of the data between entities. There should be mechanisms in place to log and timestamp the usage of data in a machine readable way.

Policies: Enforce laws (GDPR, etc.) and regulations such as purpose limitation, data minimisation, subject access rights. Policies should be machine interpretable, easily comprehensible, extensive and not disputable.

Policy generation: Allows to define policies. For instance, a user may define a rule (i.e, a policy) that forbids the creation of any preference allowing the use of medical data for advertisement.

Policy exchange: Allow entities to dynamically exchange security policy information (e.g. authentication requirements).

Manageability: Provide the appropriate tools for managing security mechanisms and policies.

Preferences: Allow for users to express preferences about the handling of their data. In particular, it must be possible to express preferences for the use of given information for a given purpose.

Autonomy: Allows users to have access to and control over their preferences without compromising the overall system.

Ownership: Allows users to get an overview, request or perform changes and deletion of their data.

Consent: Demands user consent to any changes regarding the handling of their data.

Identification: Associates each entity with a unique identifier.

Authentication: Verifies the identity of an entity to an operation or request.

Authorisation: Allows for controlling access to information based on authorisation policies attached to each entity.

Integrity: Ensures that recipient entities may detect unauthorised changes made to their data.

Confidentiality: Ensures non-disclosure of data traveling over the network to unauthorised entities.

Delegation: Allows transfer of privileges (e.g. access rights) between entities.

Anonymity/Unlinkability: Give the possibility to access data in an anonymous or pseudonymous way. However, provenance is an example for a possible conflict between transparency and unlinkability.

Traceability/Transparency: Give information on what type of data, actual value used, for what purpose and to whom the information was sent. How data is collected; how, when, where it is stored.

Completeness: Collecting complete provenance information can fully take the advance to track data and actions for identity management, error detection, etc. Incomplete provenance information may lead to detection missing and suppression of abnormal behaviors.

Scalability: With the increase of the data volume and the number of operations, it should be possible to store complete provenance information without risks of information loss.

Granularity: Not only the process derivation of a data file should be traced, but also the components of files such as paragraphs, shapes and images should be traced with regard to their origins. In short, fine-grained provenance information helps achieve highly precise anomaly detection and auditing.

Interoperability: By definition - the capability to communicate, execute programs, or transfer data among various functional units in a manner that requires entities to have little or no knowledge of the unique characteristics of those units.

Trust: Ensures trust with exchange of credentials, statement and certification, signatures, transparency and fulfilment of the other requirements.

Usability: Provides clear interfaces and structures that display security aspects, required data, digital traces, policies, possible threats in an understandable way. Usage of icons, graphs, etc. Also managing security (and privacy) is not the primary task of the user.