Technical Disclosure Commons

Defensive Publications Series

October 2019

Output-Dependent Access Control

Brett Krueger

Output-Dependent Access Control

Organizations may collect and store data, some of which may be sensitive, in data sources/stores. These organizations may have policies to safeguard sensitive data against unauthorized access or misappropriation. Traditionally, access to data sources has been controlled through the use of white lists, black lists, a combination of white lists and black lists, or other forms of Access Control Lists (ACLs). Access controls are typically based on user identifiers (IDs) or group IDs, and identify the role or function of the user or group of users. Access controls in general grant a particular actor access to particular data through a particular tool at a particular point in time. The level of granularity of data access may be at the database level, known as Coarse Grained Access Control (CGAC). But not all data may be equally sensitive. Therefore, Fine Grained Access Control (FGAC) is often used, i.e. at the level of particular tables or records within the database, or particular fields within the database tables, or even particular combinations of fields within a table. The level of access granted may depend on sensitivity of the data. Access controls may be more restrictive with respect to data containing private, privileged, or otherwise more sensitive data. Access controls may also be based on previous activity by the same actor, e.g. restricting how often the actor may run queries.

However, access control lists must be actively maintained, e.g. as individual roles change over time, or the sensitivity of data changes over time, a white list governing access to the data must be actively updated to reflect the access rights of the individual. In some cases, technology can help. For instance, privileges can be automatically revoked when employees change role within an organization, or when an employee leaves the organization, or for other reasons.

Effectively controlling access to sensitive data remains challenging, especially for large and rapidly evolving organizations and one beneficial approach is an access control approach

that collectively includes combinations of differential privacy[1], policy-based access controls, and dynamic query analysis.

Dynamic query analysis can determine the nature of the data accessed and the nature of the data contained in the results. In particular, dynamic query analysis can determine what sensitive data, if any, the query requires access to, and what sensitive data, if any, the query results disclose. Access to sensitive data may be controlled through traditional means or through policy-based access controls. That is, a policy engine may layer together several access control policies or limitations established by the organization and based on the data sources and their contents. The policy engine may only allow access to sensitive data if the user has sufficient privilege based on the aggregate access-control policies. Dynamic query analysis can determine whether traditional or policy-based access controls applies to the results of the query as well, e.g. whether the results contain sensitive information. In addition, dynamic query analysis can determine whether differential privacy should be applied to the results in order to combat data exfiltration. Differential privacy operates by adding some amount of randomness, or noise, to the results of the query in order to obscure sensitive information.

[1] https://en.wikipedia.org/wiki/Differential_privacy

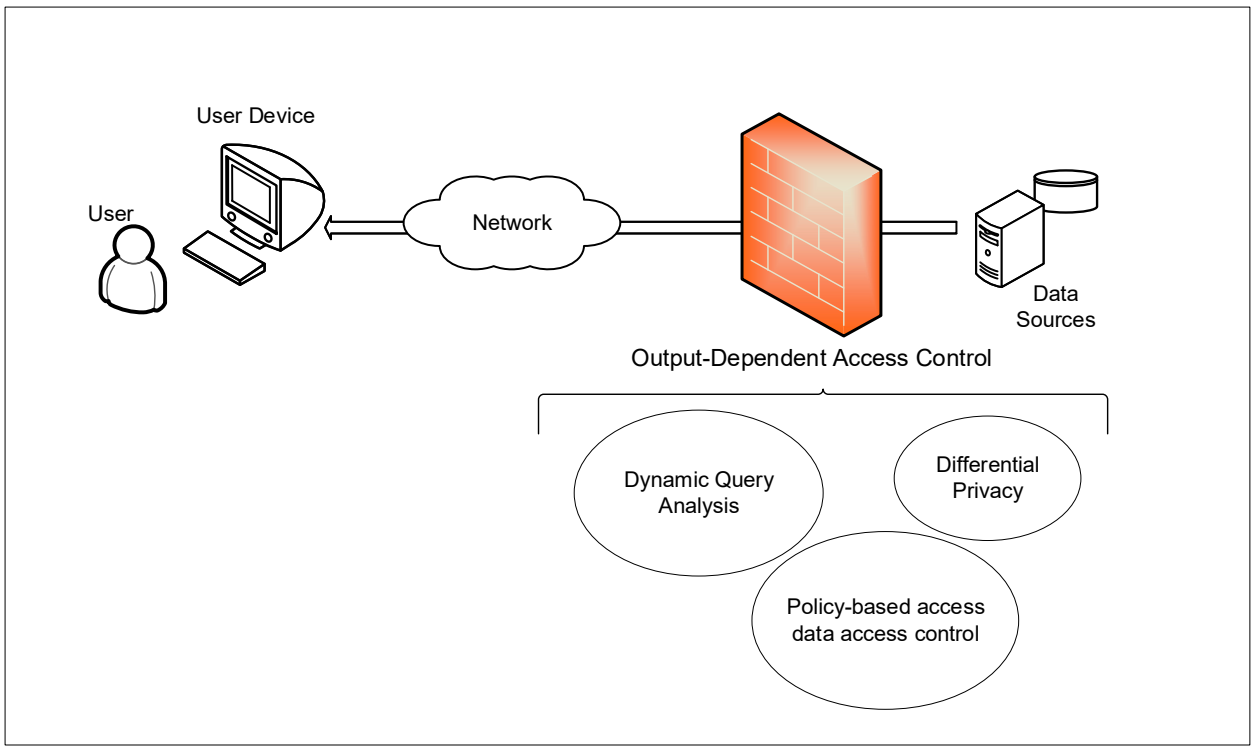

Figure 1

This process can be analogized to Erwin Schrödinger's famous thought experiment[2] applying the Copenhagen interpretation of quantum mechanics applied to everyday objects. A hypothetical cat that may be simultaneously both alive and dead is contained within a box, and its state isn't evaluated (or known) until the box is opened. In this analogy, the cat is the data and associated policies, and the box is the engine hidden from the actor's view. An actor may execute a query on a datasource and the visibility of the result is not evaluated until the final output is known or generated and compared against the policies of the upstream data.

[2] https://en.wikipedia.org/wiki/Schr%C3%B6dinger%27s_cat

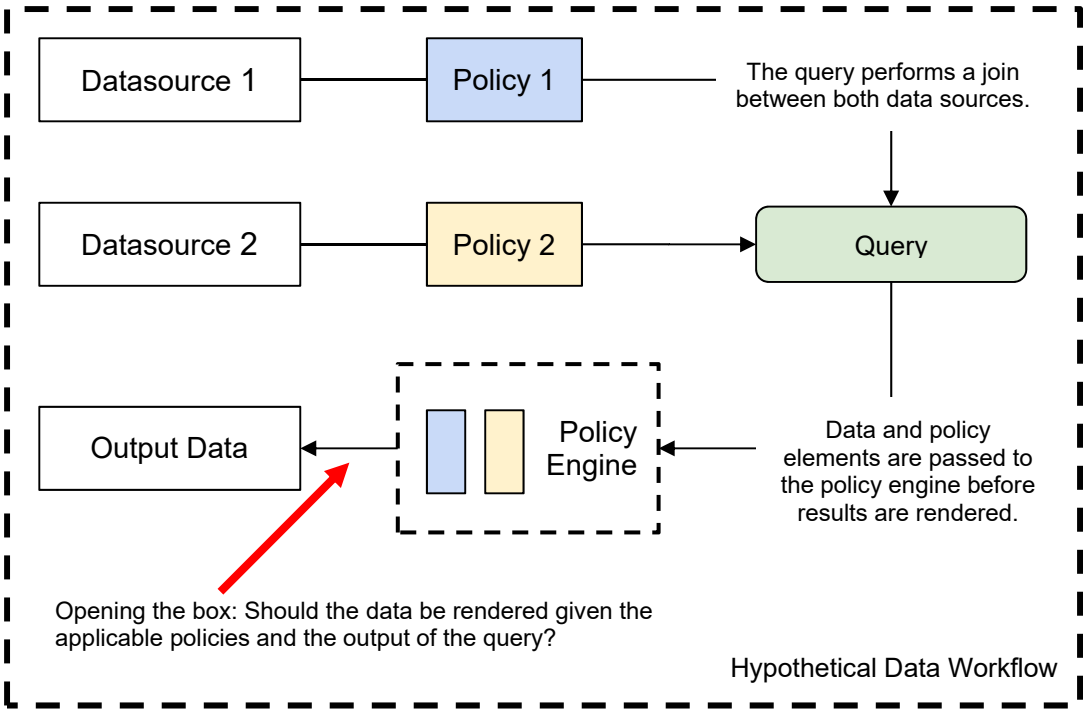

Figure 2

The state of the cat cannot be directly observed. The state of the cat is analogous to sensitive information contained within a data source, whereby the box represents the data source. Certain queries, such as "how many cats are in the box" (e.g., low-sensitive result) can be performed without revealing the state (e.g., high-sensitive result) of the cat. In other cases, differential privacy may be applied to query results to prevent revealing the state of the cat. In still other cases, the user may be privileged to know the state of the cat. In this latter case, differential privacy need not be applied.

For example, support a hypothetical employee at an organization seeks to determine the effect of a specific campaign regarding the environmental risks to drinking water from plastic disposable water bottles. In this example, one data source may contain information linking playback of a specific informational video about plastic water bottle environmental risks to the account of all employees who viewed the informational video. Another data source may contain

information linking location information of vending machines vending plastic water bottles to a campus of the organization vending the plastic water bottles to employees. A third data source may contain the purchase history of each employee that requested a plastic water bottle from the vending machines (e.g., the user may enter a user ID to the vending machine for purchasing any given refreshment (e.g., snacks, juice, water bottles, etc.), including whether the consumer purchased the plastic water bottle from the vending machine. Using traditional access control methods, the hypothetical employee would need access to all three data source, including access to the identify of each employee that viewed the informational video and purchased a plastic water bottle (or did not purchase a plastic water bottle) from one of the vending machines, as well as the location of the vending machine which one could inherently determine the location of the employee. These traditional techniques risk exposing this highly-sensitive identity information and location information of specific employees, when the end result in-fact only a low-sensitive output. E.g., "how many employees saw the informational video for the specific campaign about environmental dangers of plastic water bottles, and how many of those employees still purchased (or did not purchase) plastic water bottles from vending machines at the organization?"

Using the approach described herein, the hypothetical employee could ask "how many employees viewed the informational video and subsequently purchased refreshments from a vending machine of the organization, but did not subsequently purchase the plastic water bottle?" Dynamic query analysis would determine that the query requires access to sensitive data (subject to a privacy policy), but that the result of the query would not contain sensitive data. Therefore, the query may be permitted if the hypothetical employee has not exceeded the threshold frequency for such queries. However, dynamic query analysis may also determine that

differential privacy techniques must be applied to the results. Alternatively, the hypothetical employee could ask "which employees viewed the informational video and were subsequently purchasing refreshments from vending machines of the organization, but did not subsequently purchase the plastic water bottle?" In this case, dynamic query analysis would determine that the query output contains sensitive data, the employee identity, subject to a privacy policy. The associated policy engine would apply the appropriate policy, based on the data sources and contents, to determine what privilege is required to access the data. The policy engine may determine that the user query should be blocked, based on the user's privilege. In the first scenario, the hypothetical employee was able to uncover a valuable correlation, albeit with some degree of noise applied to preserve privacy. In the second case, the hypothetical employee was blocked. And in both cases, sensitive data was protected.

ABSTRACT

Access control management techniques can reduce the risk of data exfiltration while making privacy-trivial data insights, such as statistical correlations, more accessible. Controlling access to sensitive or private data without unduly restricting essential activities is challenging. Traditional access control techniques, including whitelists, blacklists, and Access Control Lists (ACLs) limit access to sensitive data even when the query output does not contain sensitive information. This approach uses a combination of differential privacy, policy-based access controls, and dynamic query analysis. Dynamic query analysis determines what access controls apply to the output of query, i.e. what output-dependent access controls are appropriate. When output-dependent access controls are appropriate, a policy-based engine determines what level of privilege is required. If the user lacks the required privilege, differential privacy may be applied to the results to prevent exfiltration of sensitive information.