White Paper

Protect LLMs with CipherTrust Transparent Encryption

thalestct.com

THALES
Building a future we can all trust

THALES

Introduction

The exponential growth of data in digital environments has brought about an urgent need for robust protection strategies, particularly regarding sensitive Large Language Models (LLM) use cases. How can Thales help your organization protect its private data in LLM use cases? Thales, the leader in data security solutions, offers an efficient approach to safeguarding private data in LLM use cases. Thales proposes two scenarios: Scenario 1 focuses on data protection at rest and in transit; Scenario 2 protects data at rest, in transit and during execution.

In summary, to "Protect LLMs, the Thales CipherTrust Data Security Platform with Transparent Encryption" is used, whereas organizations can leverage Thales' advanced data protection features within the CipherTrust platform.

The primary purpose is to safeguard the backend framework, which stores all data queried by users to the LLM, along with user credentials, logs, metadata, and more. Any prompt and response used and stored can contain sensitive data that requires protection. An approach to improve the efficiency of the data retrieval and accuracy of the responses using domain specific data, is known as RAG (Retrieval-Augmented Generation). Thales provides additional protection for LLMs using RAG with Thales' CipherTrust Transparent Encryption to provide a seamless and robust security framework for LLM use cases.

Ensuring data safety in LLM use cases with RAGDB requires careful security considerations. While RAG enhances query understanding and information retrieval, it also poses challenges for maintaining data privacy and security. Enhanced query understanding risks exposing sensitive data if not properly secured, and better information retrieval capabilities necessitate robust access controls and encryption to prevent unintended data access.

By addressing these risks and implementing the proposed solution, organizations can effectively fortify their data protection strategies and mitigate the risks associated with LLM use cases, thereby ensuring the confidentiality and integrity of sensitive information.

Why to protect data at rest for LLM's

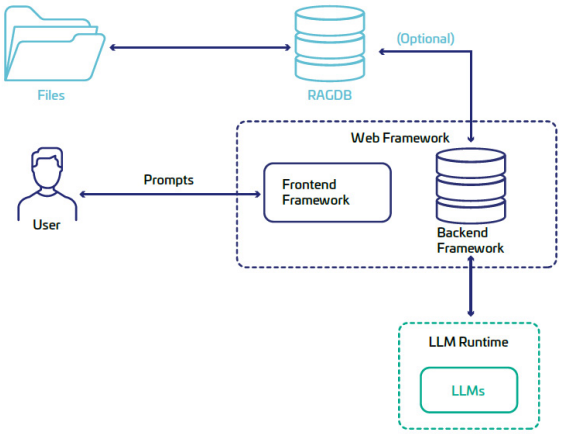
Applications without LLMs function according to predefined rules and logic, making their outputs deterministic. This means that, as long as the rules remain unchanged, the same inputs will consistently produce the same outputs. This deterministic nature simplifies the process of validating and testing rules to ensure they perform correctly in all intended scenarios. Additionally, it is generally straightforward to trace which rules lead to specific actions and to understand their application.

LLMs, on the other hand, operate fundamentally differently. An LLM has no hard-coded rules; only neurons in layers that are connected by weights and biases. In this context, a neuron is a computational unit within a layer of an LLM that receives embeddings as inputs, applies a mathematical function to those inputs, and produces an output that is passed to the next layer of neurons. These connections between neurons affect how the inputs influence the probability of generating an output token as the inputs pass through the layers.[1]

Additionally, the models themselves do not provide any trust boundaries. There is no differentiation between "code" and "data" when using an LLM, and LLMs can not provide fine-grained access control to information contained within the model's weights. Assume that with the ability to interact with the LLM, a sufficiently motivated adversary can retrieve any data that the LLM was trained upon (Nasr et al., 2023).

A comprehensive understanding of these models and the security implications posed by a non-deterministic component within the broader system is essential. Due to their fundamentally probabilistic nature, LLMs create new opportunities for malicious users to exploit access to sensitive data or manipulate the models into taking unintended actions. This can lead to severe consequences, such as compromised model integrity, that can lead to data breaches. It can also result in non-compliance with privacy, security and AI-specific regulations.

For these reasons, the LLM should be considered an untrusted entity in the context of other, more deterministic traditional components.



Background

This figure shows the basic components and architecture for interacting with custom-built LLMs. The RAGDB (with the files) are optional components depending on whether the user is using prompt stuffing techniques or utilizing the capabilities of a RAG. The knowledge base in the form of files and documents are retrieved by the RAGDB. The web framework, specifically the backend of the web framework, usually handles all the communication between all the components. Note that, depending on the specific framework, the backend framework (within the web framework) might also function as a separate, independent component dedicated solely to interacting with all other components, including the web framework itself. The LLM runtime will also communicate via its APIs with the backend framework. The user will interact with the front-end framework where the UI will display the conversation (between the user and the LLMs). In the following sections, we will present two scenarios that use CipherTrust Data Security Platform (CDSP) components to protect LLMs. First, we will describe the CDSP components and then describe how they secure the architecture illustrated here.

Scenario 1: Data protection at rest and in transit

Components

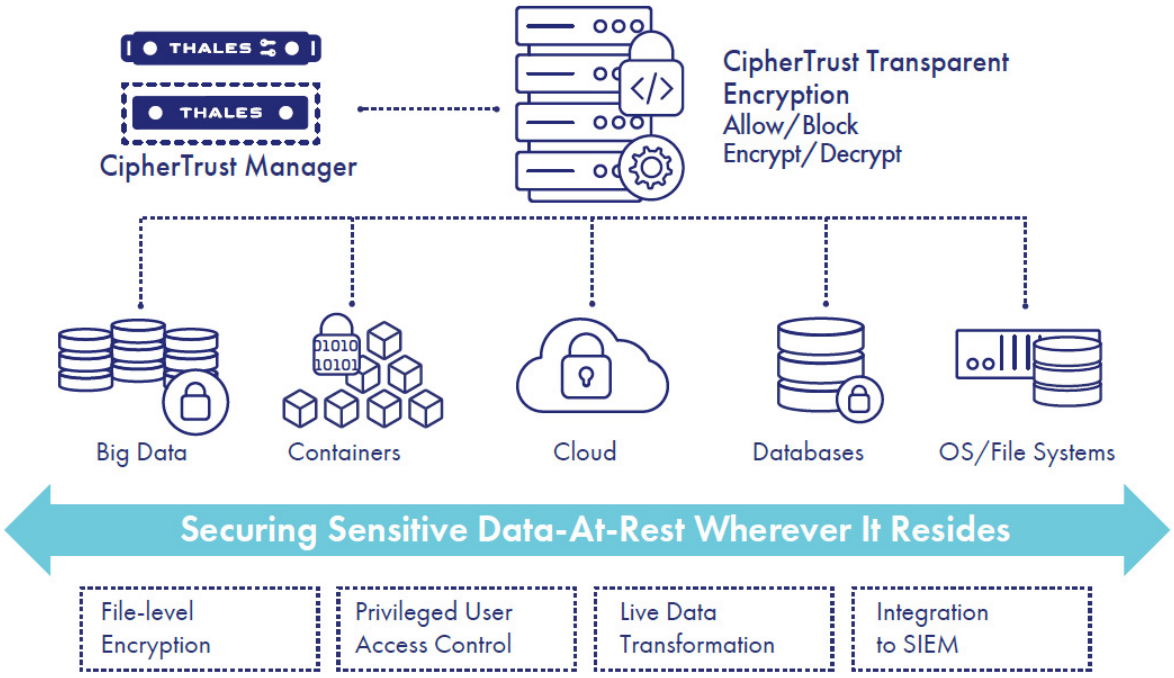
CipherTrust Manager

CipherTrust Manager (CM) is a part of the CipherTrust Data Security Platform (CDSP). CDSP is a data-centric solution that significantly reduces risk across your business and decreases the number of resources required to maintain strong data security. CDSP integrates centralized data security and management core functions, such as key management, data discovery and classification, data protection and granular access controls. By centralizing and simplifying data security, CDSP provides an efficient integrated fast data protection platform, centrally managed by the customer, that helps accelerate the time to compliance and safe cloud migrations of sensitive data. IT teams ask for a data-centric solution that secures data as it moves from networks to applications and the cloud. When perimeter network controls and endpoint security measures fail, data-centric solutions enable organizations to remain compliant with evolving privacy regulations and the demand to support a tremendous number of remote employees. CM can be deployed on premises, in cloud or hybrid environments, or subscribed to as a service.

As the central management point for CDSP, CM simplifies key lifecycle management tasks for all your encryption keys. This industry leading enterprise key management solution manages secure key generation, backup/restore, clustering, deactivation, deletion, and access to Connectors and partner integrations that support a variety of use cases (e.g., data discovery, data-at-rest encryption, enterprise key management, and cloud key management). CM supports role-based access control to keys and policies, robust auditing, and reporting, and offers development- and management-friendly REST APIs. CM is available in both physical and virtual form factors. Hardware and virtual appliances can leverage embedded Luna Network HSMs (Hardware Security Module) or select cloud HSMs to enable FIPS 140-2 Level 3 highest level root of trust.

CipherTrust Transparent Encryption

CipherTrust Transparent Encryption (CTE), is part of the CDSP, delivers data-at-rest encryption with centralized key management, privileged user access control and detailed data access audit logging. Protecting data wherever it resides, on-premises, across multiple clouds and within big data, and Kubernetes environments. The deployment is simple, scalable, and fast, with agents installed at operating, filesystem or device layer, and encryption and decryption are transparent to all applications that run above it. CipherTrust Transparent Encryption is designed to meet data security compliance and best practice requirements with minimal disruption, effort, and cost. Implementation is seamless keeping both business and operational processes working without changes even during deployment and roll out. The solution works in conjunction with the FIPS 140 up to Level 3 compliant CipherTrust Manager, which centralizes encryption key and policy management for the CipherTrust Data Security Platform.



LLM Framework

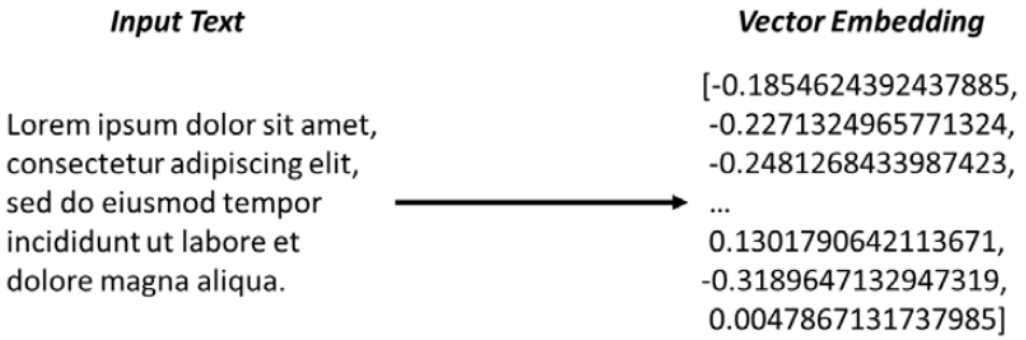
A major component in using any LLM will be its LLM framework or runtime. An LLM framework will be a platform or tool focused on simplifying the deployment and use of large language models (LLMs). The main goal is to provide a user-friendly interface and infrastructure to facilitate the integration of advanced AI (Artificial Intelligence) capabilities into various applications. Key features to look for in an LLM framework would typically include:

- Ease of Use: Simplifying the interaction with LLMs through intuitive interfaces or APIs.
- Scalability: Ensuring that the platform can handle large-scale deployments and heavy computational loads.
- Flexibility: Supporting a range of models and customization options to cater to different use cases.
- Integration: Providing seamless integration with existing tools and workflows to enhance productivity and efficiency.

RAGDB

Before getting into the details about RAGDB, it is important to understand RAG itself. Retrieval-Augmented Generation (RAG) is a hybrid approach that combines retrieval-based methods with generation-based methods to improve the quality and accuracy of responses generated by language models. It leverages an external knowledge base or database to provide more contextually relevant information during the generation process. It has 3 main components described below:

- **Embedding Models Component:**
  - **Function:** An embedding model is a machine learning algorithm that trains to represent information as dense representations in a multi-dimensional space.
  - **Mechanism:** The process, called embedding, converts high-dimensional data into low-dimensional vectors, which simplifies the data and makes it easier for machine learning algorithms to process. They are models that are trained specifically to generate vector embeddings, which are long arrays of numbers that represent semantic meaning for a given sequence of text, as shown in the image below. The resulting vector embedding arrays can then be stored in a database, which will compare them to search for similar data.



- **Retrieval Component:**
  - **Function:** This component searches a large corpus of documents (knowledge base) to find relevant information related to a given query.
  - **Mechanism:** It typically uses a dense retrieval model, like a bi-encoder architecture where both the query and documents are encoded into dense vectors. The retrieval process involves finding documents whose vectors are closest to the query vector in the embedding space.
- **Generation Component:**
  - **Function:** This component generates coherent and contextually relevant text based on the information retrieved by the retrieval component.
  - **Mechanism:** It uses a sequence-to-sequence generation model, often based on transformers like GPT or BERT, which can produce fluent and contextually appropriate text.

RAGDB (Retrieval-Augmented Generation with Databases) is an extension of the RAG framework by integrating structured databases into the retrieval-augmented generation process. This approach enhances the ability to produce accurate and contextually relevant text by utilizing structured and often more precise information from databases. Even though RAGDB has the same 2 components as RAG, there are some slight differences within the retrieval component. RAGDB includes Structured Retrieval; In addition to unstructured text, RAGDB retrieves data from structured databases. This involves querying relational databases, knowledge graphs, or other structured data sources to find specific information.

Web Framework

The web or application framework refers to the software structure and tools used to develop the interface and logic that users interact with. These are open platforms or libraries for operating large language models (LLMs) in production. It also enables developers to easily run inference with any open-source LLMs, deploy to the cloud or on-premises, and build powerful AI apps. As most web applications, a web framework should contain the fronted and backend components like mentioned below

Frontend

- **User Interface (UI):** Developed using web technologies like HTML, CSS, and JavaScript (often with frameworks like React or Vue.js) to create the interactive elements of the web application.
- **User Input Handling:** Captures user questions and displays the responses from the LLM.

Backend

- **API Gateway:** Manages requests from the front-end and forwards them to the LLM services.
- **Authentication:** Ensures secure access to the model services.
- **Session Management:** Maintains context for ongoing user interactions.
- **Request Processing:** Handles preprocessing of user inputs before sending them to the LLM.
- **Response Formatting:** Post-processes the model outputs for display on the frontend.

CTE Transparent Encryption for LLM Architecture

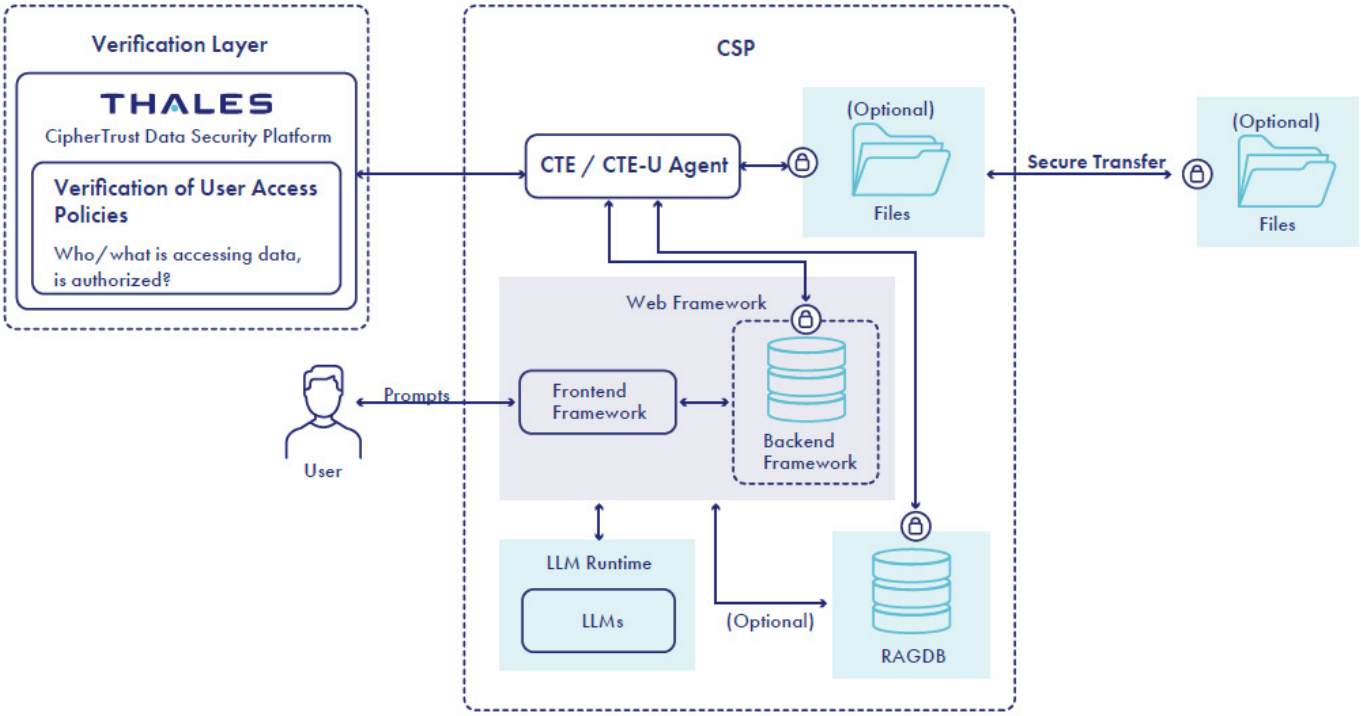
The proposed solution provides a general architecture framework for protecting the LLMs with CipherTrust Transparent Encryption (CTE). The components mentioned above play a crucial role in establishing a robust protection mechanism for protecting sensitive data. The RAGDB components are labelled optional as the user can perform the LLM tasks without utilizing its functionality but, as it is proven to be a powerful way to specialize LLM responses to a specific domain corpus, it remains as an important option if the user wishes to enhance their use case. Every time the file system storage is accessed in this solution, it is mapped as GuardPoints which are protected through the CTE policies.

In the diagram below, there are 2 main sections - the verification layer and the components inside the Cloud Service Provider (CSP) along with a user who is interacting with those components. The CTE / CTE-U Agent is the heart of the architecture within the CSP.

- **CTE agent:** The CTE (CipherTrust Transparent Encryption) agent is a software component installed at the kernel level on a physical or virtual machine. It protects data on that machine. Once installed, the CTE agent enables encryption/decryption for any number of devices or directories on the machine, ensuring that only authorized users and processes can access the encrypted data at file system level transparent to the application.
- **CTE-U agent:** The CTE-U (CipherTrust Transparent Encryption User Space) agent is a user-space level encryption solution that serves the same fundamental purpose as the CTE agent. Unlike the CTE agent, the CTE-U agent operates at the user-space level.

The CTE / CTE-U agent is responsible for the GuardPoints of the web framework, the uploaded files and the RAGDB. A GuardPoint specifies the list of folders that contain paths to be protected. The access to the files and the encryption/decryption of those files under the GuardPoint is controlled by security policies. The lock and key symbol in the diagram depict a GuardPoint folder. The CM is used in the verification layer to attest the user policies which are set by the authorized user. The CM checks who or what is accessing the data and whether they are authorized. The agent also handles all the encryption and decryption of the GuardPoints (with the key management and attestation of the CM).

The Files component in the diagram refers to the knowledge base required by the LLM to generate relevant outputs. These files need to be deployed through an existing guardpoint with matching policies, enabling the files to remain encrypted during transit. The files are retrieved by the RAGDB. Within the RAGDB, the embedding models transform the data into vector embeddings. These embeddings are then sent to the LLM, via the LLM runtime and backend web framework. The LLM uses the vector embeddings to generate the relevant response, and it sends back a response to the RAGDB. The RAGDB will transform the LLM vector embedding response back to user readable text. This response will then be sent back to the web framework where it can be shown to the user via the UI (using a frontend framework).



Benefits

- **Protect data at rest:** All the stored data (data at rest) in the whole architecture is protected using the GuardPoints. This ensures that the sensitive data does not inadvertently be exposed by the RAGDB or from within the LLM environment
- **Access Control protection:** CipherTrust Manager prevents unauthorized exposure of sensitive data using its robust access control using access policies.
- **Enhanced data protection:** Stringent data protection strategies can be implemented by use of flexible but reliable polices set by the authorized user to ensure that sensitive information is not inadvertently included in generated responses

Scenario 2: Data protection at rest, in transit, and during execution

Thales also provides cutting edge confidential computing solutions for cloud infrastructure to protect sensitive data all the time: At rest, in transit and in use with End-To-End Data Protection (E2EDP). E2EDP is supported by CM on various cloud service providers, using Confidential Computing (CC) technologies, trusted cloud independent attestation and verification. E2EDP is based on the principle of separation of duties, where the customer remains in control of its own data protection for cloud deployments and defines the profile of cloud hardware/software stack where its workloads will be computed. This approach enhances trust in cloud deployments by holding each stakeholder responsible for their respective roles and reduces the ability for a malicious actor to access code and data at rest, in transit and while being executed. The customers can migrate existing workloads with sensitive data or create new workloads needing zero trust, confidential computing, and Confidential AI to broaden data security, attestation and set the right authorizations. With end-to-end data protection, multiple parties can securely collaborate on various use cases, such as Confidential

AI datasets and models as needed while preserving privacy, confidentiality, and compliance with privacy regulations.

In addition to the components of Scenario 1, Scenario 2 incorporates the following components as described in the next section.

Confidential VM

- Confidential VMs (Virtual Machines) or CVM (Confidential VMs) are a type of virtual machine designed to ensure the confidentiality and integrity of data and applications while they are in use. This concept is particularly important in cloud computing environments where multiple tenants might share the same physical hardware. Traditionally, data has been protected at rest and in transit, but with confidential computing technologies, this protection is extended to data in use. Confidential computing protects data in use by performing computations in a cryptographically isolated hardware-based Trusted Execution Environment (TEE). Here are some key features and benefits of confidential VMs:
  - Data Confidentiality: Confidential VMs use hardware-based encryption to protect the data in memory. This means that even if an attacker gains access to the physical server, they cannot read the data stored in the VM's memory.
  - Hardware Based Trusted Execution Environment (TEE): Leveraging silicon level encryption, the enclave isolates the computation from the rest of the system, ensuring that only authorized code can access the data.
  - Integrity: Confidential VMs ensure that the code running in the VM has not been tampered with. They achieve this by verifying the integrity of the software stack, including the operating system and applications.
  - Regulatory Compliance: By ensuring data confidentiality and integrity, confidential VMs help organizations comply with regulatory requirements for data protection, such as GDPR (General Data Protection Regulation), HIPAA, and others.
  - Secure Multi-Tenancy: Confidential VMs enable secure multi-tenancy by isolating each tenant's data and computations even from the CSP host.

3rd Party Attestation Service

A Zero Trust 3rd party attestation service is required to verify the trustworthiness of the computation assets at the network, edge, and in the cloud. This service attests the cloud hardware and software stack as defined by the customer policies. Leveraging the Confidential Computer hardware's root of trust, the service verifies that the customer workloads in the cloud operate in a legitimate Trusted Execution Environment as expected. Specifically, this service would check the hardware and software policies of the cloud infrastructure, and whether these measurements are aligned with the security policies defined by the customer. The attestation process provides cloud independent assurance to any relying party that the TEE, any data, and workloads running within it have not been compromised.

Root of Trust during CPU and GPU Inference

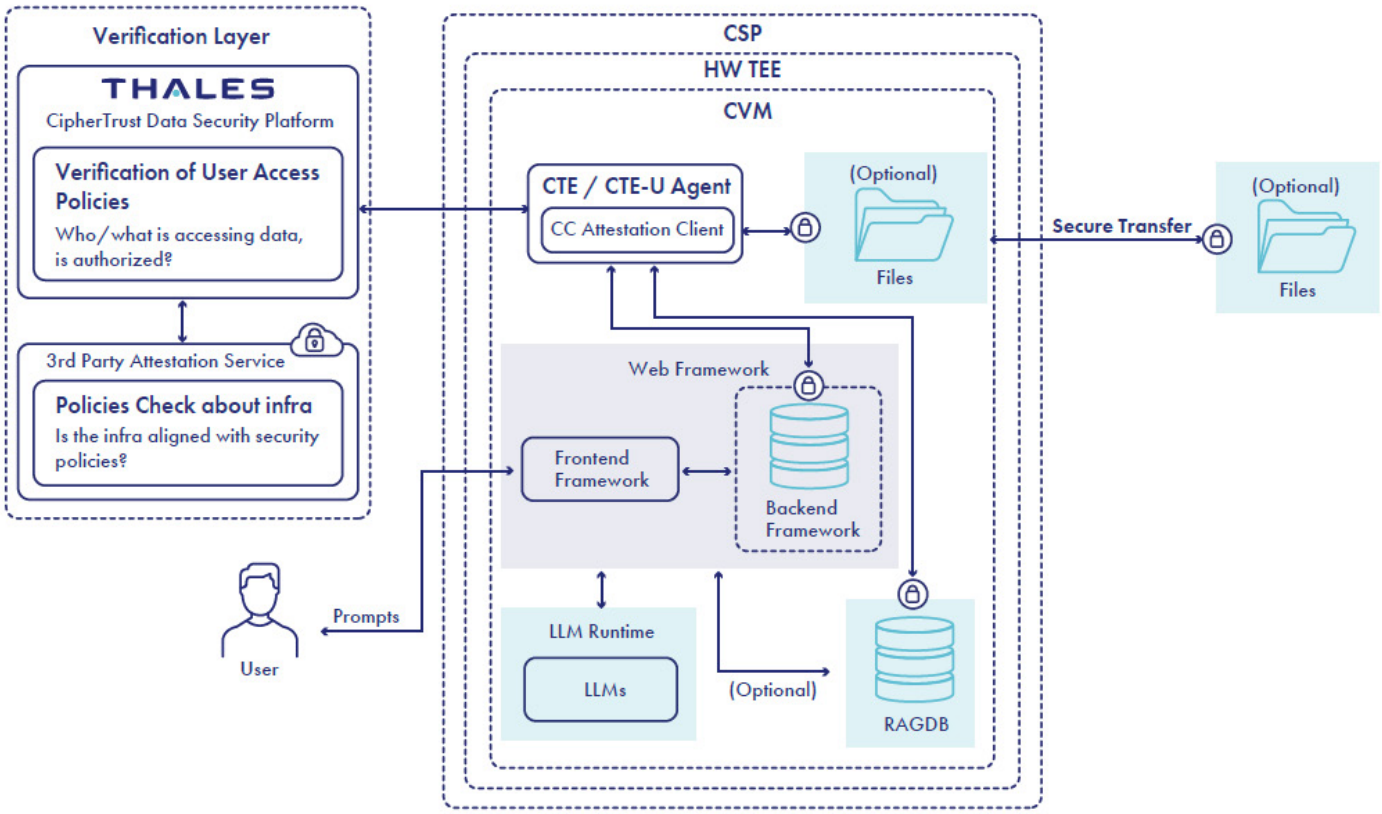
There are extra hardware dependencies which need to be taken into consideration when using a confidential computing solution. This refers to the difference in the root of trust for a CPU and a GPU when used in the E2EDP solution. If CPUs are used for the LLM inference, then the CPU inference runs within the confidential VM and thus it is within the root of trust of the enclave (TEE). If GPUs are used for the LLM inference, then the GPU inference runs within the runtime memory of the GPU, which is not part of a chain of trust of the enclave and requires its own attestation and verification.

The E2EDP architecture

To enable E2EDP, the generic solution provided explained previously for CTE is complemented with 3rd party attestation service, the confidential computing environment itself and verification.

Thales's CM can integrate 3rd party attestation services seamlessly. The CTE/CTE-U agent is enriched with an embedded CC attestation client for confidential computing use cases. The CTE/CTE-U agent will send the information/measurements regarding the infrastructure of the confidential environment to the 3rd party attestation service, via the CM. The 3rd party attestation service will check the polices set for the given infrastructure and will attest if the infrastructure information aligns with the customer policies.

The confidential computing environment consists of a Hardware Trusted Execution environment (HW TEE) and a CVM. As mentioned before, hardware-based TEE is a secure area in the main processor, which helps the code and data loaded inside it be protected with respect to confidentiality and integrity. The only difference with the generic solution is containing all the components of the CSP within the CVM, to make sure that only the authorized user can view the sensitive information (in the context of LLMs).



Additional Benefits

- **Runtime memory protection:** E2EDP solution provides runtime at memory protection, for a complete and holistic data at execution protection solution. Data comes from an encrypted at rest GuardPoint into a protected memory space.
- **Separation of duties:** This solution is fundamentally based on separation of duties using hardware-based separation of the underlying software, admins, and other cloud tenants. The duties for "Key Provider", "Cloud Provider" and "Attestation Provider" are separated into individual and independent roles.

Conclusion

In today's data-driven landscape, safeguarding sensitive information in LLM use cases is paramount. Thales can help your organization protect private sensitive data in LLM use cases as explained herein with Thales CipherTrust Transparent Encryption. Whether your organization is leveraging the advanced capabilities of Retrieval-Augmented Generation (RAG) or not, combined with state-of-the-art data protection, CTE provides a robust framework to protect private data in different scenarios as needed by the use case.

Thales' solution enhances the protection for query understanding, information retrieval, and contextual responses by addressing the challenges these advantages pose to data privacy and security. By implementing stringent data protection strategies, such as robust access controls and transparent encryption, organizations can mitigate the risks associated with LLM use cases. Offering advanced data security features with cutting-edge confidential computing technologies End-to-end data protection is the ultimate solution for data protection at rest, in transit and during execution, guarantees the confidentiality and integrity of sensitive information. This approach effectively strengthens data protection strategies in an era of exponential data growth.

About Thales Trusted Cyber Technologies

Thales Trusted Cyber Technologies, a business area of Thales Defense & Security, Inc., is a trusted, U.S. provider of cybersecurity solutions dedicated to U.S. Government. We protect the government's most vital data from the core to the cloud to the edge with a unified approach to data protection. Our solutions reduce the risks associated with the most critical attack vectors and address the government's most stringent encryption, key management, and access control requirements.

For more information, visit www.thalestct.com