Encrypted Search and Data-Centric Security with Virtru

Introduction

Secure data sharing has become essential for organizations across all sectors. As sensitive information is increasingly exchanged over email, collaboration tools, and cloud platforms, robust encryption is crucial to ensure data remains confidential. Traditional security models based on perimeter defenses are no longer sufficient, especially in today's cloud-native, distributed environments. Data-centric security — where protection is embedded into the data itself — offers a more comprehensive approach to safeguarding information in transit and at rest.

However, as organizations encrypt their data, they face a common challenge: searching and retrieving specific pieces of encrypted information without compromising security. This paper delves into Virtru's approach to encrypted search, detailing the technologies and methodologies that enable organizations to balance strong encryption with the ability to efficiently search encrypted data for legal, compliance, or operational purposes.

The Challenge: Balancing Security with Accessibility

Organizations handling sensitive data must adhere to strict security protocols, including data encryption at rest and in transit. However, encrypting data often comes at the cost of usability, particularly regarding search functionality. For example, legal teams conducting audits or compliance officers executing e-discovery processes must search emails and documents efficiently, even when encrypted.

Historically, enterprises have faced a tradeoff: maintaining robust encryption while sacrificing the ability to search encrypted content. Without searchable encryption, audits, compliance checks, and e-discovery become cumbersome and ineffective. Virtru solves this dilemma with its patented encrypted search technology, which enables secure keyword searches on encrypted content without decrypting the data.

Virtru's Approach: Searchable Encryption

Virtru addresses this challenge with its patented Searchable Symmetric Encryption technology. This approach allows organizations to search through encrypted data without exposing the plaintext search terms to unauthorized parties, including the service provider.

- **Tokenization and subtoken expansion:** Virtru's solution converts each word in an encrypted email or document into a four-character alphanumeric token. For example, "confidential" may become "a0b1". Each token is further divided into smaller subtokens to prevent correlation attacks, and random tokens, or noise, are added. This approach makes it nearly impossible to deduce the original word from the tokenized data, enhancing privacy.
- **HMAC-based search tokenization:** When a user initiates a search, the search terms are transformed using a Hash-based Message Authentication Code (HMAC). The search terms are never exposed, as the transformation process ensures that the original input remains secure. The user's query is matched against the encrypted tokens without decrypting the content, maintaining full confidentiality during search operations. Moreover, organizations will use different keys for the HMAC conversion so that organization A won't have the same tokens per word as organizations B or C.
- **Bloom filters:** A key component of Virtru's searchable encryption system is its use of Bloom filters — a probabilistic data structure that enables secure and efficient search operations. By storing tokenized data within Bloom filters, Virtru enables searches without maintaining the original word order or exposing plaintext content. The structure obfuscates the relationship between the encrypted content and search terms, adding a layer of complexity that prevents reverse engineering.
- **Noise injection:** To further obfuscate search tokens and prevent any statistical analysis that could reveal sensitive data, Virtru introduces random noise tokens into the tokenized data. This technique ensures that attackers cannot reliably determine the original data based on repeated patterns, even if attackers gain access to the encrypted tokens.

Example: Probabilistically Searchable Email Generation

To help understand these key technologies, refer to the diagram below, which explains the step-by-step process of transforming plaintext email content into searchable, encrypted tokens.

1. Unecrypted HTML	" Hello, World! "
2. Normalize to plaintext	"Hello, World!"
3. Tokenize text to words	["hello","world"]
4. HMAC each word	["2F10D3812A2C6E2F…","9A2C6E2F2F10D381…"]
5. Map to 3x indices of chosen size. For example: 4-digit Base36 [000-zzz]{36^3}=0-46655	[[z1x3, 7ubw, 020b], [v5c1, 7ubw, xz9z]]
6. Store in Bloom filter (removes order and redundancy)	
7. If needed, random noise entries are included until a false-positive rate is reached	[020b, v5c1, 7ubw, g9rz, pb21, xz9z, z1x3]
8. Turn Bloom filter entries back into 4-letter words and store in email metadata	"020b v5c1 7ubw g9rz pv21 xz9z z1x3" "hello" → "(z1x3 AND 7ubw AND 020b)" "hello" → "(z1x3 AND 7ubw AND 020b)" OR Hello

Here's a breakdown of each step:

1. **Unencrypted HTML:** The email starts in its raw HTML format, such as " Hello, World! ". At this stage, the content is in its unencrypted form.

2. **Normalize to plaintext:** The HTML structure is stripped away, leaving only the actual text, such as "Hello, World!". This simplifies the processing of content, as it removes any non-text elements.

3. **Tokenize text to words:** The normalized text is split into individual words, e.g., ["hello", "world"]. This step is crucial for applying cryptographic operations to each word.

4. **HMAC each word:** A Hash-based Message Authentication Code (HMAC) is applied to each word. For example, "hello" is transformed into a hashed string like "2F10D3812A2C6E2F...", and "world" into another hashed string. The HMAC ensures the word is encrypted securely but can still be searched later without exposing the plaintext.

5. **Map to indices**: The HMAC result for each word is mapped to four indices using Base36 encoding (ranging from 0000 to zzzz, representing numbers from 0 to 46655). For example, the word "hello" might be mapped to [z1x3, 7ubw, 060b], while "world" could be mapped to [5c1d, 7u9w, az9z]. These indices serve as probabilistic representations of the words.

6. **Store in Bloom filter:** The indices are stored in a Bloom filter, a probabilistic data structure used to test whether an element is part of a set. The Bloom filter removes the order and redundancy of the indices, making it impossible to deduce the original text sequence or content.

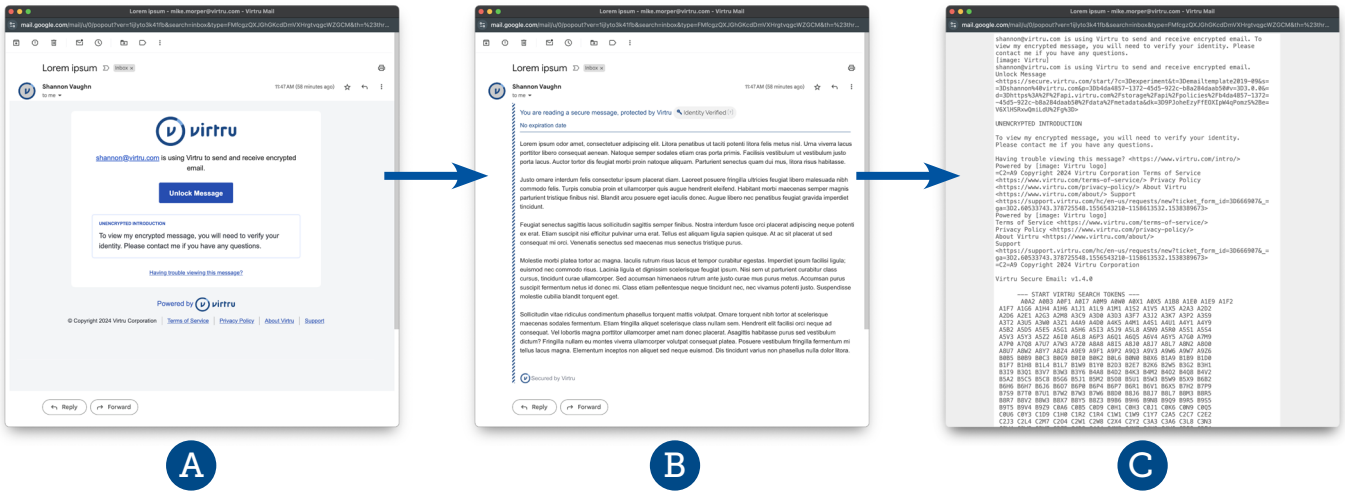
7. **Random entries for False Positive Rate:** To maintain a controlled false-positive rate (FPR), additional random entries are added to the Bloom filter if necessary. This ensures that searches might return occasional false positives, which enhances security by preventing patterns from revealing the actual content.

8. **Convert Bloom filter entries back:** The indices from the Bloom filter are turned back into four-letter or digit combinations, which are stored in the metadata of the email, e.g., "0s0b 5yc1 7u8w 9rz bi21 zp9z 2zx3". These indices represent the encrypted form of the words. When a user searches for the term "hello," it is matched with its probabilistic representation like "(z1x3 AND 7ubw AND 060b)", allowing the system to find emails that contain the encrypted version of "hello" without revealing the actual word.

This process enables searchable encryption by converting plaintext into hashed tokens stored in a Bloom filter. The tokens are randomized and expanded to prevent correlation attacks, allowing users to search encrypted emails securely without revealing the plaintext content.

When a search query is made, the corresponding tokens for the search term are generated and compared against the stored tokens. The system retrieves the email or document matching those tokens without revealing the original search term or data. This method guarantees that even during a search, the privacy of the content is maintained, and no plaintext is exposed at any point during the process.

Example: How Search Tokens Are Embedded in Virtru Protected Email Messages



A. **Encrypted Message:** Before the recipient's identity is confirmed, a Virtru-protected email message remains encrypted in the recipient's inbox

B. **Decrypted Message:** After the recipient's identity has been confirmed, a Virtru-protected email message is decrypted and viewable

C. **Search Tokens:** By selecting "Show Source," the embedded Virtru search tokens can be viewed (highlighted in red box). These tokens are not visible in the decrypted (B) message, however they are present.

Real-World Application: Encrypted Search in Action

Virtru's encrypted search technology is deployed across industries with stringent data protection requirements, such as healthcare, finance, legal, and government sectors. These organizations must protect sensitive data and maintain the ability to access and search it when necessary. Some specific use cases include:

Legal and E-Discovery: Legal departments and law firms require secure access to encrypted documents during audits and litigation. Virtru's encrypted search enables them to search through encrypted emails and files while ensuring that no unauthorized party gains access to the sensitive content.

Healthcare: Healthcare organizations must comply with regulations like HIPAA, which mandates strong encryption of patient records and medical communications. Virtru's encrypted search enables healthcare providers to retrieve specific information from encrypted data sets without compromising patient confidentiality.

Financial Services: Financial institutions must safeguard sensitive customer data, such as transaction histories and personal information. Virtru's technology enables these institutions to comply with regulations like GDPR and CCPA while allowing compliance officers to perform audits and searches on encrypted data.

Leveraging Virtru's Patent for Encrypted Search

Virtru's encrypted search technology, as outlined in its patent – Methods and systems for generating probabilistically searchable messages – is a groundbreaking approach that allows users to perform secure, full-text searches on encrypted content. The patent details how tokenization and subtoken expansion work with HMAC-based token matching to enable searches without exposing the original plaintext data.

One key innovation is position-independent tokenization, where tokens representing individual words are decoupled from their position within the document, thus eliminating the risk of revealing word patterns through token analysis. Additionally, the split-knowledge architecture ensures that search keys and content never reside in the same location, minimizing the risk of unauthorized access.

Benefits of Virtru's Encrypted Search

- **Preservation of Security and Privacy:** Virtru's encrypted search technology ensures that the search process remains confidential. Neither the data owner nor the service provider has access to the search terms or the decrypted content, protecting the privacy of sensitive data.

- **Seamless Integration with Existing Workflows:** Virtru's encrypted search integrates with major platforms like Google Drive, Microsoft 365, Gmail, Microsoft SharePoint, and more, allowing users to search within their familiar environments. This ease of integration lowers organizations' adoption barriers.

- **Compliance and Audit Capabilities:** Virtru's solution is tailored to meet the compliance needs of industries subject to strict data protection regulations. Encrypted search ensures that organizations meet audit and e-discovery requirements without decrypting sensitive data.

- **Zero Trust Architecture:** By implementing Virtru's split-knowledge encrypted search, organizations adopt a Zero Trust model, where no entity—including the service provider—has full access to the data. This architecture reduces the risks of unauthorized access, even if perimeter defenses are compromised.

Conclusion: Empowering Secure, Searchable Data

Virtru's patented encrypted search technology represents a significant advancement in data-centric security. By enabling keyword searches on fully encrypted content, Virtru allows organizations to protect sensitive data without sacrificing functionality. This balance between security and usability is crucial in industries that rely on compliance, audits, and secure collaboration.

With Virtru's encrypted search, organizations can confidently implement strong encryption while maintaining the ability to efficiently search for specific information, ensuring both data protection and operational efficiency.

Contact us to learn more how Virtru can help your organization implement data-centric security.