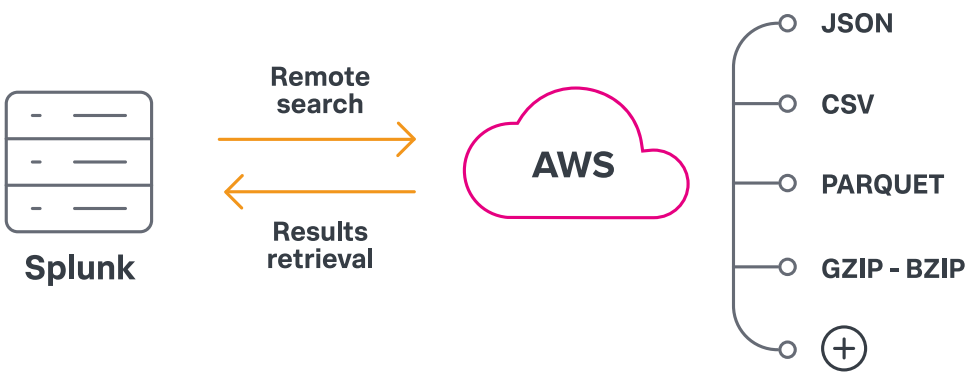
Federated Search for S3 Tech Brief

Technology landscapes have become increasingly more complex, which means the world of data is constantly changing. While data continues to grow exponentially, not all data is created equal — data has different formats and quality and changes with age. Additionally, cost restrictions can lead to data silos, which inhibit holistic visibility and result in inefficient use. This motivates organizations to adopt cloud object storage services to address their needs related to scalability, performance, cost efficiency, security and compliance. Splunk can now search Amazon S3, the largest cloud object storage service in the market today.

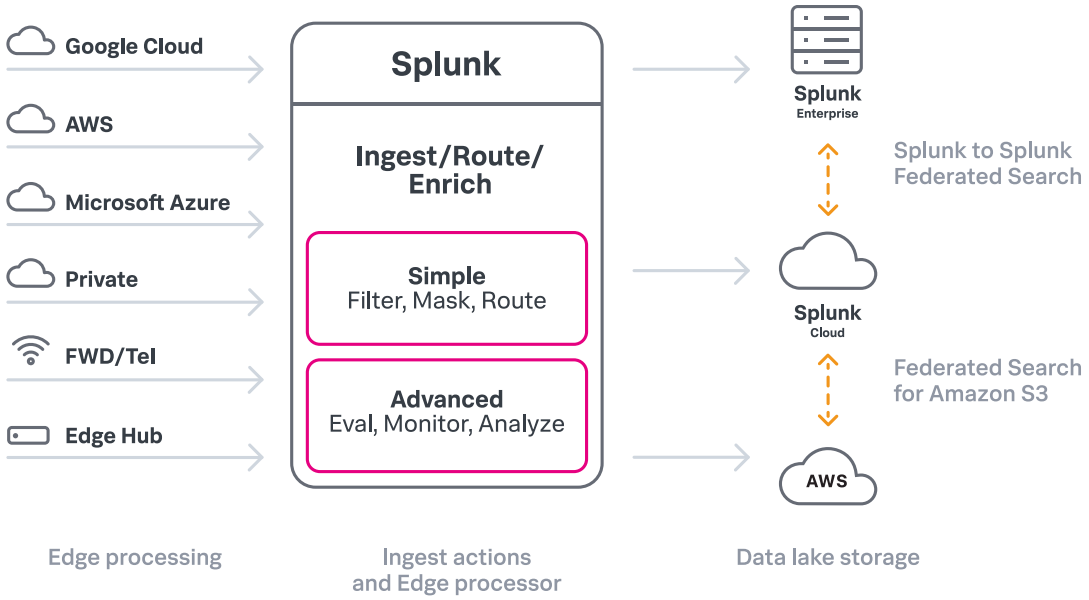
What is Federated Search for Amazon S3?

Federated Search for Amazon S3 is a remote search capability that allows Splunk users to use SPL to search their Amazon S3 buckets at rest and retrieve the results directly in their Splunk Cloud Platform stack without ingesting the data. This allows them to search, invest, enrich, analyze and correlate data between Amazon S3 and Splunk.



How does this fit with the rest of your Splunk environment?

Splunk allows you to bring together data from across your organization's hybrid, multi-cloud technology landscape. Splunk allows you to ingest the data regardless of the source, with features like **data manager** which allows you to easily onboard new data sources. This is ideal for data that needs real-time search latency and high frequency access patterns. Then, with **ingest actions** and **edge processor** you can easily filter, mask, enrich and route your data to Amazon S3. Federated Search for Amazon S3 will fulfill the need for accessing historical data at rest for ad-hoc searching, as well as scenarios like security investigations and ITOps analytics. For example, as you're onboarding a verbose data source, you may want to retain some of the high-value data in Splunk and route the rest to Amazon S3 to retain for historical purposes. If you then need to search the data sent to Amazon S3, or read new data that was created directly in Amazon S3, you can use **Federated Search for Amazon S3** to seamlessly search your Amazon S3 buckets (currently available only on Splunk Cloud Platform). If you have multiple Splunk environments, **federated search** allows you to execute a unified search across multiple Splunk environments (Splunk Cloud Platform and Splunk Enterprise).



Use Cases for Federated Search for Amazon S3

Federated Search for Amazon S3 provides a convenient search mechanism to access historical datasets on Amazon S3 without ingesting data. This can be leveraged to empower these search use cases:

1. **Security investigations over historical data:** Security users can run infrequent investigation searches over data in Amazon S3 which may contain multiple years worth of logs. This use case is an improvement of a process that would typically require data rehydration on an ad-hoc basis, which is a lengthy and expensive process that impacts customers' time-to-detection.
2. **Statistical analysis over historical data:** Splunk users can run infrequent statistical analysis over data in Amazon S3 that may contain multiple years worth of logs to obtain key insights based on properties of their logs. Customers would typically use external tools to run statistics or aggregate searches.
3. **Data enrichment**: Splunk users can enrich their existing data with additional context and information by creating lookups from data in S3. When trying to understand more about your data, manually looking up additional information in Amazon S3 takes time. You can enrich Splunk indexed events and automatically add information from Amazon S3, accelerating the overall process.
4. **Data exploration:** Customers would typically use external tools to explore their Amazon S3 datasets, or take a conservative approach with new datasets to minimize Splunk costs. Splunk admins can now explore new datasets stored on Amazon S3 in order to determine the usefulness of the data. After this analysis, customers can plan to ingest business-critical subsets of the data from Amazon S3 into Splunk.
5. **Archive scale management:** Customers need to keep a historical record of data for compliance purposes. They want to achieve that using Amazon S3 while still leveraging Splunk search in case investigations are needed. Customers can use DDAA to manage their Splunk Cloud Platform data, but they are conscious of costs and delays when restoring if they ever need to search that data.
6. **Manage infrequently accessed data:** Customers try to store as much data as they can in case they need to search it due to regulation, compliance or an investigation. However, this often results in storage of redundant data or datasets that are not frequently searched.

Federated Search for Amazon S3 is not intended for high frequency and real-time searches. You **should not** replace your existing search scenarios involving high frequency searches or real-time latency requirements. Using this feature for real-time searches will result in slower performance and reduced search functionality when compared with indexed Splunk searches on ingested data. Using this feature for high frequency searches is likely going to result in higher costs than natively ingesting/searching in Splunk.

Getting started

To get started with federated search of structured Amazon S3 data, you need:

- A Splunk Cloud Platform deployment on AWS with Federated Search for Amazon S3 enabled
- Federated Search for Amazon S3 Data Scan Units *(Note: This is an add-on paid feature)*
- An AWS account with data in Amazon S3 buckets that conforms to supported file and compression types
- One or more AWS Glue Data Catalog tables

Learn more

24-350877 -Splunk Federated Search with AWS S3 Tech Brief -102