PUT YOUR DATA SHARING IN MOTION

Why data streaming should be part of your data sharing strategy

Comprehensive data is critical for government missions to succeed. The sources of this data are not confined to organizational boundaries. For the complex decision making required by government agencies, data must extend beyond bureau, component, and agency boundaries. Agencies must seek to share data to the consuming organizations that need it when they need it. Data sharing covers myriad use cases, missions, and domains with data consumers desiring access in these ways:

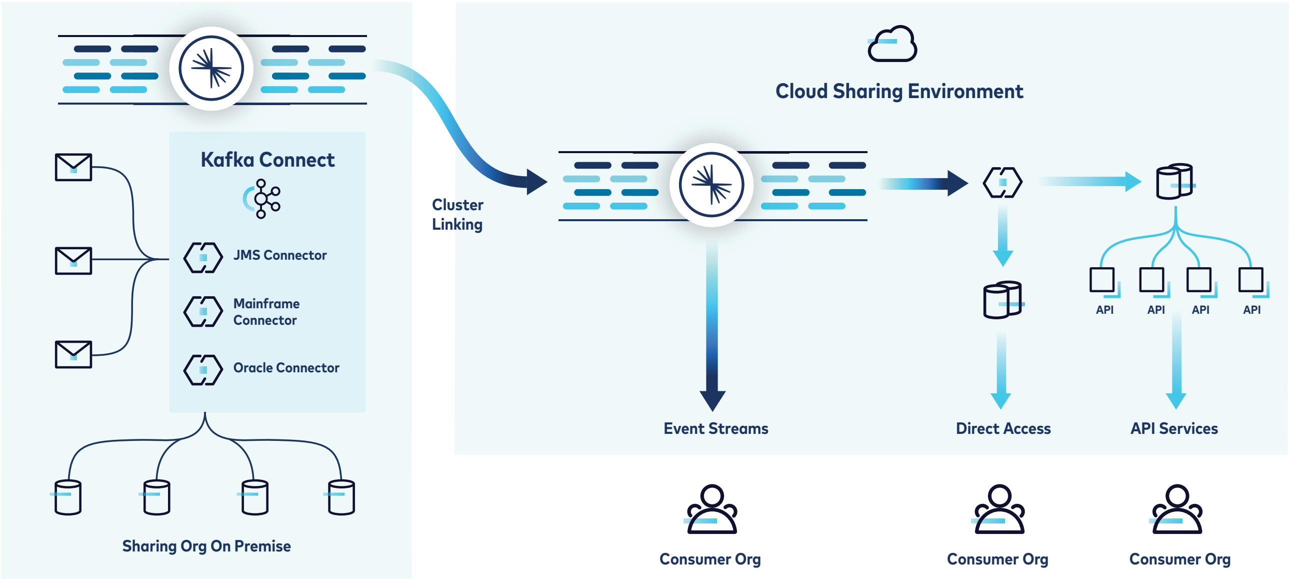
Scattered requests The consumer needs a specific record or a very small subset of a larger body of data in an ad hoc fashion and does not have ultra sensitive time requirements. This could be checking the processing status of an application or a background check for criminal records. Scattered requests can often be the results of a user action.

Event driven The consumer wants a constant feed of events or updates from a data source so they can efficiently react to data as it is created. Examples could be the DoD acting on aircraft data from the FAA or other sources, law enforcement acting on vehicles crossing the border, or public healthcare responding to infections.

Synchronization of data In this scenario the consumer wants to maintain its own separate version of the data set. The largest driver of this approach is that the source can not effectively support the sort of queries and analysis that is required. In this case the data is copied into a local workload specific store such as a data lake or graph database. Another reason for synchronization is the need to have higher uptimes or lower latency responses than the connecting network or source can guarantee. Security requirements or differences are yet another frequent motivator. This data sharing pattern is the most common need in government today as agencies are looking to make data-based decisions across many bodies of data.

Batch access On an ad hoc or infrequent basis consumers need to grab a specific batch of data for analysis and then will likely throw it away. Examples of this are things like scientific analysis of relatively non-dynamic data, research on historical data, etc.

Architectural view of how Confluent can enable the three main modalities of data sharing from legacy on-premise data sources up through a cloud environment



Access pattern	API	Data streaming	Direct access	Download
Scattered Record Access	●	○	◑	○
Data Synchronization	◔	●	◔	○
Event Driven	◔	●	◔	○
Ad hoc Batch	◔	◑	◑	●

Common data sharing approaches

With government agencies understanding the power and value of data to drive more actionable and impactful decisions, agencies are still working through the most efficient way to derive that value and meet the needs of the scenarios listed above. To satisfy data sharing demands a few common approaches have been employed.

Web Service API Service oriented APIs, and in particular RESTful services, are well understood, easy to use, and have a rich ecosystem of tooling for both development and operations. This approach is very effective for scattered requests but awkward and inefficient for event driven consumption patterns or synchronization.

Direct download With direct download, an entire set of data is extracted, frequently compressed, then made available for consumers to download using a common protocol like HTTP or FTP. This is a simple and long standing way of sharing data, and is often used to support science and research use cases where having easy access to an entire dataset is more important than ensuring that data is up to date. This method is only really appropriate for a subset of the batch access demands.

Direct access With this technique, data owners provide direct access to data stored in their infrastructure. Until recently, this was rarely adopted due to concerns around security and the impact to the operational systems. This is changing as data sharing specific capabilities within data stores has improved and an increased adoption of elastic cloud capabilities. This method can be useful for scattered requests and batch access but is not effective for synchronization or event driven data sharing. It's important for the data provider to be aware of the costs that can easily add up from continuous processing and egress of the same data over and over again. For example, if a consumer analyzes the entire data set in their own tool on an hourly basis then the same data will be transferred 24 times a day at a huge cost to the data producer.

While each of these common data sharing approaches are good for certain data sharing needs, none of them are particularly well suited for data synchronization or the event driven data sharing that is needed to efficiently power fast and real time applications.

Modern data sharing

Data streaming enables organizations to put data sharing in motion. The sharing organization publishes a stream of events (including changes and deltas) as they occur and data sharing consumers can subscribe to efficiently receive them as they happen. Since data steaming is built from the ground up for the event driven approach, it is perfectly aligned to event driven sharing. In addition, the nature of sharing the changes or deltas to data is a fundamental principle of data replication, making it ideal for data synchronization.

Confluent, powered by the de facto standard in data streaming, Apache Kafka®, is built and optimized to support the distribution of event streams whether that is within a project, across an organization, or between organizational entities. Confluent runs anywhere from the edge, to the data center, to the cloud. By providing a massive ecosystem of connectors, organizations can use them to extract events and changes from their operations data stores, modern or legacy, and curate them for inter organizational consumption. Coupled with cluster linking, Confluent provides an easy path to move data from on prem into your cloud environment for data sharing.

Confluent provides the ability for data consumers to act immediately on events as they are received without requiring a copy to be made. This enables organizations to have an up-to-date copy of source data in a technology of choice.

- Data streaming gives agencies:
- A continuous feed of the events
- A continuous and efficient data synchronization by just sending deltas
- Ability to conduct event-driven action
- Access to high volume data processing

Data sharing with Confluent means that data consumers can react immediately as data is created at the source. Simultaneously, costs are optimized by only sharing changes and using a technology designed for sharing data with multiple down stream consumers.

To learn more, please visit www.confluent.io.

Ready to get started? Contact a Confluent expert today

Email us on publicsector@confluent.io
Or visit confluent.io/get-started for more details