Data Provider to Application Provider Interface

Purpose: The purpose of this interface is for the Application Provider to discover, request, and receive data from the Data Provider.

Interface Description:

The interface is generally initiated by the application provider and typically consists of the following types of interactions.

1. Authentication/Authorization - Most data providers will require at least some form of authentication/authorization from the connecting client. This authentication may be a simple as the provider simply validating the IP address and name of the client or may be as complex as a PKI level certification. While a few data providers may allow anonymous access at a minimum they will track the connection source and data transferred. A number standards for authentication and authorization exist today and range from community accepted best practices full formal standards endorsed by national bodies and SDOs.
2. Discovery – A discovery request by the Application Provider will retrieve information about the data available from the Data Provider. This request may be as simple as a directory listing (when the data provider is a file server) or a more complex request in the form of a query about database schema. Typically, the application provider is looking to obtain metadata about the available data on the provider. Meta data may be as simple as file names and sizes or as complex as complete schemas and associated semantic information (ontologies). In addition, the discovery request may not be handled directly by the data provider but rather is handled by a separate metadata catalog service. In many cases this service may require a completely separate connection and authentication/authorization from the one to the data provider. Standards for both the metadata specifications and metadata discovery are developing. Most metadata standards are functional domain specific and lay out the mandatory and optional metadata required for that specific data type or data domain. Discovery standards are also evolving and provide for standardized representations of data attributes and semantics. In some cases such as ontology information the discovery standard simply references a data representation standard. It should be noted that the discovery request itself may result in a large data transfer.
3. Data Request – The request for data can take many forms ranging from a simple request for a specific data item (e.g. retrieve this file), to a request for all data on a topic (e.g. pub/sub), to a simple query (give me all records where a data value is X), to a complex request involving complex data transformation (e.g. relation joins, summarization, aggregation, etc.). It is in this case where the software transfer depicted in the RA would take place. Software in this case would again cover a wide range from a simple SQL query string to a full blown remote batch job submission. The key reason that software would be required is that in the case of big data the raw/bulk data would be too large to transfer and some reformatting, thinning, or mining is required to make the transfer size reasonable. In situations where the Data Provider is itself a big data system the request may be a full blown analytic that performs very complex analysis on a large data set and returns a smaller but still large data set.
4. Data Transfer – The data transfer interaction can also take many forms and may involve synchronous or asynchronous communications. In the simple form the appropriate connection is made (generally a network connection) from the data provider to the application provider and the data is passed as a single chunk. This would generally be like a file transfer. In cases where the Application provider cannot accept all of the data in bulk the data may be transferred in blocks/chunks based in a synchronous manner. Finally, the data may be sent from the provider as a continuous stream or in the case of a pub/sub or event processing type request asynchronously as individual records/objects as the provider obtains new data.
5. Data Storage – As the data is transferred the application provider needs to store the data someplace. For the purposes of this discussion in-memory persistence of the data will be treated the same as long term persistence. In general, the application provider will interact with the framework provider to store/persist the data as required.

Roles within the Big Data Application Provider

Within the Big Data Application Provider the interactions described above may be accomplished as shown in the table below by the following application provider roles.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Big Data Application Provider Roles** | | | | |
| **Interaction** | **Collection** | **Preparation** | **Analytics** | **Visualization** | **Access** |
| Authentication/Authorization | Primary |  |  |  |  |
| Discovery | Secondary | Primary | Secondary |  |  |
| Data Request | Secondary | Primary | Secondary |  |  |
| Data Transfer | Primary | Secondary | Secondary |  |  |
| Data Storage | Secondary | Primary | Secondary |  |  |