Loading Parquet Content

Overview

A Parquet file resembles a text file of comma-separated values such as one might export from Excel. It is a preferred file format in an HDFS system.

The Attivio Intelligence Engine (AIE) provides a tool for ingesting Parquet content.

View incoming links. **>>**

No links were found.

Parquet Scanner

A **ParquetScanner** parses a Parquet file and sends its contents to AIE as a set of IngestDocuments . The ParquetScanner produces one IngestDocument for each line of the Parquet file.

Configuring the AIE Schema

The columns of a Parquet file will be mapped into fields in the IngestDocument. The field names can be automatically stripped from the top line of the file, if that is appropriate, or they can be specified in the connector configuration. Either way, the field names must be registered in the AIE Schema or they will not be admitted to the index.

See Configure the Attivio Schema for further information.

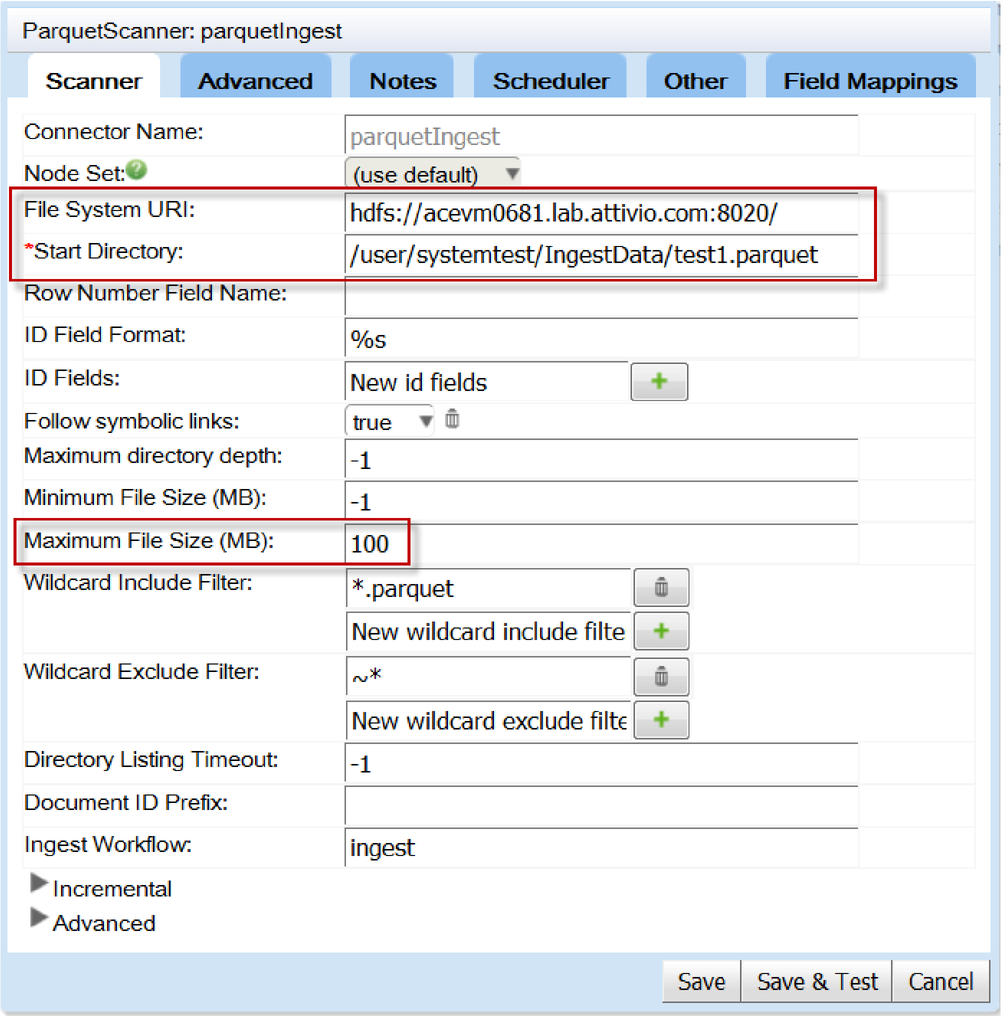
Configuring a Parquet Connector

You can configure a Parquet connector by using the Connector UI.

Start AIE using the Command-Line Interface. This will start AIE and will make the Administration UI available at **http://<host>:17000/admin**. Note that the AIE Agent must be running before you attempt to start AIE.

In the AIE Administrator, navigate to **System Management > Connectors**. Click **New** in the menu bar. Select the **Parquet Files** connector from the list.

On the **Scanner** tab of the resulting dialog box, enter the **Connector Name**, the **File System URI**, and the **Start Directory**. Check the **Maximum File Size** setting, which could easily be too low.



ParquetScanner: parquetIngest

Connector Name:	parquetIngest
Node Set:	(use default)
File System URI:	hdfs://acevm0681.lab.attivio.com:8020/
*Start Directory:	/user/systemtest/IngestData/test1.parquet
Row Number Field Name:	
ID Field Format:	%s
ID Fields:	New id fields
Follow symbolic links:	true
Maximum directory depth:	-1
Minimum File Size (MB):	-1
Maximum File Size (MB):	100
Wildcard Include Filter:	*.parquet New wildcard include filte
Wildcard Exclude Filter:	~* New wildcard exclude filte
Directory Listing Timeout:	-1
Document ID Prefix:	
Ingest Workflow:	ingest

▶ Incremental

▶ Advanced

Click **Save.** The Connector Editor writes out the connection configuration to the project's configuration servers. Note that you will eventually need to **Update** your project from the AIE-CLI in order to copy the connector's configuration file to the projects sources.

Parquet Connector Properties

The ParquetScanner is configured by setting properties on the editor.

Parquet Scanner Tab	Remarks
Connector Name	The name of the connector as seen in the UI or in XML.
Node Set	The nodeset the connector should run on. Defaults to default-service-nodeset. The Editor can set this value only on new, unsaved connectors.
File System URI	Use this field to access an HDFS file system. The syntax is hdfs://[username@] host:port, for example, hdfs://acevm0681.lab.attivio.com:8020/. Otherwise leave it empty. [REQUIRED].
Start Directory	The directory containing the files to scan, or the root directory of the tree to scan.[REQUIRED]. Avoid using the same start directory in multiple Parquet scanners. This can confuse the incremental deletion feature, causing unexpected deletions.
Row Number Field Name	Name of the field to put the line number in.
ID Field Format	Describes how to concatenate the values from one or more **idFields** into a single value, which will be used as the record's unique id. The value is a string that follows the behavior of the **format** method of the Java String class.

ID Fields	A list of Parquet fields to concatenate to create a unique id value. Used with idFieldFormat. Default is "id".
Follow Symbolic Links	Whether or not the scanner should follow symbolic links while crawling the file system.
Maximum Directory Depth	Maximum number of nested directory levels to traverse. "-1" means no limit.
Minimum File Size (MB)	Minimum file size to send (in MB). Smaller files will be dropped.
Maximum File Size (MB)	Maximum file size to send in megabytes.
Wildcard Include Filter	File-extension wildcards. Matching files will be scanned.
Wildcard Exclude Filter	File-extension wildcards. Matching files will not be scanned.
Directory Listing Timeout	Provide configurable directory listing times (in seconds).
Document ID Prefix	Append this prefix to the Document ID during processing.
Ingest Workflow	Ingestion workflow to receive the ingested documents. String.
Incremental	
Incremental Mode Activated	Enables incremental updates. Boolean.
Incremental Deletes	Optional. Used with 'incremental-activated' parameter to control if AIE should delete documents that have been removed from the source files. Default is true.
Advanced	
Delete After Crawl	Boolean. Delete the files after they have been scanned. Do not use with the incrementalModeActivated feature.
Move to directory after crawl	Move the scanned files to this directory after they are scanned. Do not use with the incrementalModeActivated feature.
Additional Start Directories	If there is *only one* root directory to scan, put it in the **Start Directory** field and optionally specify a **Move to Directory After Crawl** directory where the files should be placed after the crawl. If there is more than one root directory to scan, put the *first one* in the **Start Directory** field (and optionally specify the **Move to Directory After Crawl** field) and then add the other directories here. Each entry is two strings. The first string is the Start Directory. The second string is the optional Move To Directory After Crawl directory.
Max Rows	Number of rows to read from the file.

The "other" tab contains Kerberos settings:

Parquet Other Tab	Remarks
Keytab	Location of keytab file for Kerberos authentication.
Principal Name	Principal name for Kerberos authorization.
Name Node Principal	Configuration property for enabling support for Kerberos.
Scan hidden files	If true, scan all readable files including system and hidden files.

The other tabs in the Connector Editor are described on the Connectors page.

Running the ParquetConnector

Erasing the Index

While testing a new connector, you will frequently need to empty the index and try again. Methods of deleting the index are described here.

To run the ParquetConnector, open the AIE Administration UI, and navigate to the **System Management > Connectors** page. Right-click on **ParquetConnector** and click on **Start**.

Then navigate to SAIL, which is **Query > SAIL.** Search for *:*, which retrieves all records in all tables. We can see that the scanner was successful:

To view all of these fields in the search results, open the **SAIL Properties** dialog box (click on the gear icon) and add the field names to the **Field Expressions** tab **Other Results Fields** list.

Incremental Updating

The primary role of incremental updating is to load new Parquet content while skipping over files that have already been ingested. If a Parquet file is deleted after the first run of an incrementally-enabled Parquet connector, the next run deletes all of the documents that came from that file. If the file is left in place with one or more rows removed, however, the next connector run does *not* delete documents associated with the missing rows, because it only checks whether the source file is still present.

This connector supports the Activating Incremental Updating features. There is a tutorial example of incremental updating here.

⚠ After running the connector to ingest documents with Incremental Mode activated, be careful with any future configuration changes to the connector, as such changes can cause one or more of the following issues:

- Some incremental changes might not be properly identified, and hence, not get ingested into AIE in future runs.
- Some documents can remain in your index that are no longer managed by any connector. These documents can eventually become out of date and contain outdated content security permissions.

If you must make changes to change the connector configuration after running it, follow these steps to keep your system fully up to date:
1. Delete any previous documents the connector created in your AIE index.
2. Select your connector from the AIE Administrator's Connectors tab, and Reset the connector.