# RDB via Snapshots Introduction

## Features

The RDB connector via Snapshots will crawl content from any relational database that can be accessed using JDBC. The connector will extract data based on SQL statements and submit this data in to Aspire for processing.  The connector is different from many other connectors as it directly extracts the data, so typically there's not a *fetch data* phase. However, if your database includes references to external data (say URLs to web sites or paths of external files), then a *fetch*stage may be invoked.

RDB connector via Snapshots features include the following:

- Connects to database server using JDBC drivers (these must be downloaded separately)
- Performs full crawling
- Performs incremental crawling, so that only new/updated documents are indexed, using snapshot files
- Fetches data from the database using SQL statements
- Is search engine independent
- Runs from any machine with access to the given database

### Content Retrieved by the Connector

The content retrieved by the connector is entirely defined using SQL statements, so you can select all or subsets of columns from one or more tables. Initially, the data is inserted in to Aspire using the returned column names, but this may be changed by further Aspire processing.

## JDBC Drivers

The RDB via Snapshots connects to databases via JDBC, so you'll need the appropriate JDBC client (driver) JAR file for the database you want to connect to. These are available for most (if not all) major database vendors, and your first port of call for the driver should be the vendor's website.

## Operation Mode

### Retrieve Data per Batch

This mode uses SQL taken from the job (*<connectorSource/discoverySQL>*, *<connectorSource/extractSQL>* or configuration) and execute them against the database configured via a Multi RDBMS Connection Pool stage. Each resulting row is formed into an AspireObject using the column names as document elements, and this document is submitted to a pipeline manager using the event configured for inserts. As the document is created, the value of the column identified in the job (*<connectorSource/idColumn>*) is noted as the primary key of the document. The value *insert* will be placed in the *action* attr ibute of the document.

Column names from the extractSQL query are added to the AspireObject inside the "connectorSpecific" field. If the column names are standard AspireObje ct fields, they are added to the root level. See Connector Metadata for further details on which are standard fields.

Any change detected in the query set in *discoverySQL* field will be compare with the snapshot file and report the change if required.

### Retrieve Everything

This mode uses SQL taken from the job (*<connectorSource/fullSQL>* or configuration) and execute them against the database configured via a Multi RDBMS Connection Pool stage. Each resulting row is formed into an AspireObject using the column names as document elements, and this document is submitted to a pipeline manager using the event configured for inserts. As the document is created, the value of the column identified in the job (*<connector Source/idColumn>*) is noted as the primary key of the document. The value *insert* will be placed in the *action* attribute of the document.

Column names from SQL queries are added to the AspireObject inside the "connectorSpecific" field. If the column names are standard AspireObject fields, they are added to the root level. See Connector AspireObject Metadata for further details on which are standard fields.

Any change detected in the query set in *fullSQL* field will be compare with the snapshot file and report the change if required.