OGR SQL

The **GDALDataset** supports executing commands against a datasource via the **GDALDataset::ExecuteSQL()** method. While in theory any sort of command could be handled this way, in practice the mechanism is used to provide a subset of SQL SELECT capability to applications. This page discusses the generic SQL implementation implemented within OGR, and issue with driver specific SQL support.

Since GDAL/OGR 1.10, an alternate "dialect", the SQLite dialect, can be used instead of the OGRSQL dialect. Refer to the SQLite SQL dialect page for more details.

The **OGRLayer** class also supports applying an attribute query filter to features returned using the **OGRLayer::SetAttributeFilter()** method. The syntax for the attribute filter is the same as the WHERE clause in the OGR SQL SELECT statement. So everything here with regard to the WHERE clause applies in the context of the SetAttributeFilter() method.

NOTE: OGR SQL has been reimplemented for GDAL/OGR 1.8.0. Many features discussed below, notably arithmetic expressions, and expressions in the field list, were not support in GDAL/OGR 1.7.x and earlier. See RFC 28 for details of the new features in GDAL/OGR 1.8.0.

SELECT

The SELECT statement is used to fetch layer features (analogous to table rows in an RDBMS) with the result of the query represented as a temporary layer of features. The layers of the datasource are analogous to tables in an RDBMS and feature attributes are analogous to column values. The simplest form of OGR SQL SELECT statement looks like this:

```
SELECT * FROM polylayer
```

In this case all features are fetched from the layer named "polylayer", and all attributes of those features are returned. This is essentially equivalent to accessing the layer directly. In this example the "*" is the list of fields to fetch from the layer, with "*" meaning that all fields should be fetched.

This slightly more sophisticated form still pulls all features from the layer but the schema will only contain the EAS_ID and PROP_VALUE attributes. Any other attributes would be discarded.

```
SELECT eas_id, prop_value FROM polylayer
```

A much more ambitious SELECT, restricting the features fetched with a WHERE clause, and sorting the results might look like:

```
SELECT * from polylayer WHERE prop_value > 220000.0 ORDER BY prop_value DESC
```

This select statement will produce a table with just one feature, with one attribute (named something like "count_eas_id") containing the number of distinct values of the eas_id attribute.

```
SELECT COUNT(DISTINCT eas_id) FROM polylayer
```

General syntax

The general syntax of a SELECT statement is:

```
SELECT [fields] FROM layer_name [JOIN ...] [WHERE ...] [ORDER BY ...] [LIMIT ...] [OFFSET ...]
```

Field List Operators

The field list is a comma separate list of the fields to be carried into the output features from the source layer. They will appear on output features in the order they appear on in the field list, so the field list may be used to re-order the fields.

A special form of the field list uses the DISTINCT keyword. This returns a list of all the distinct values of the named attribute. When the DISTINCT keyword is used, only one attribute may appear in the field list. The DISTINCT keyword may be used against any type of field. Currently the distinctness test against a string value is case insensitive in OGR SQL. The result of a SELECT with a DISTINCT keyword is a layer with one column (named the same as the field operated on), and one feature per distinct value. Geometries are discarded. The distinct values are assembled in memory, so a lot of memory may be used for datasets with a large number of distinct values.

```
SELECT DISTINCT areacode FROM polylayer
```

There are also several summarization operators that may be applied to columns. When a summarization operator is applied to any field, then all fields must have summarization operators applied. The summarization operators are COUNT (a count of instances), AVG (numerical average), SUM (numerical sum), MIN (lexical or numerical minimum), and MAX (lexical or numerical maximum). This example produces a variety of summarization information on parcel property values:

```
SELECT MIN(prop_value), MAX(prop_value), AVG(prop_value), SUM(prop_value),
       COUNT(prop_value) FROM polylayer WHERE prov_name = 'Ontario'
```

It is also possible to apply the COUNT() operator to a DISTINCT SELECT to get a count of distinct values, for instance:

```
SELECT COUNT(DISTINCT areacode) FROM polylayer
```

Note: prior to OGR 1.9.0, null values were counted in COUNT(column_name) or COUNT(DISTINCT column_name), which was not conformant with the SQL standard. Since OGR 1.9.0, only non-null values are counted.

As a special case, the COUNT() operator can be given a "*" argument instead of a field name which is a short form for count all the records.

```
SELECT COUNT(*) FROM polylayer
```

Field names can also be prefixed by a table name though this is only really meaningful when performing joins. It is further demonstrated in the JOIN section.

Field definitions can also be complex expressions using arithmetic, and functional operators. However, the DISTINCT keyword, and summarization operators like MIN, MAX, AVG and SUM may not be applied to expression fields. Starting with GDAL 2.0, boolean resulting expressions (comparisons, logical operators) can also be used.

```
SELECT cost+tax from invoice
```

or

```
SELECT CONCAT(owner_first_name,' ',owner_last_name) from properties
```

Functions

Starting with OGR 1.8.2, the SUBSTR function can be used to extract a substring from a string. Its syntax is the following one : SUBSTR(string_expr, start_offset [, length]). It extracts a substring of string_expr, starting at offset start_offset (1 being the first character of string_expr, 2 the second one, etc...). If start_offset is a negative value, the substring is extracted from the end of the string (-1 is the last character of the string, -2 the character before the last character, ...). If length is specified, up to length characters are extracted from the string. Otherwise the remainder of the string is extracted.

Note: for the time being, the character as considered to be equivalent to bytes, which may not be appropriate for multi-byte encodings like UTF-8.

```
SELECT SUBSTR('abcdef',1,2) FROM xxx   --> 'ab'
SELECT SUBSTR('abcdef',4)   FROM xxx   --> 'def'
SELECT SUBSTR('abcdef',-2)  FROM xxx   --> 'ef'
```

Starting with OGR 2.0, the *hstore_get_value()* function can be used to extract a value associate to a key from a HSTORE string, formatted like 'key=>value,other_key=>other_value,...'

```
SELECT hstore_get_value('a => b, "key with space"=> "value with space"', 'key with space') FROM xxx -->
        'value with space'
```

Using the field name alias

OGR SQL supports renaming the fields following the SQL92 specification by using the AS keyword according to the following example:

```
SELECT *, OGR_STYLE AS STYLE FROM polylayer
```

The field name alias can be used as the last operation in the column specification. Therefore we cannot rename the fields inside an operator, but we can rename whole column expression, like these two:

```
SELECT COUNT(areacode) AS "count" FROM polylayer
SELECT dollars/100.0 AS cents FROM polylayer
```

Changing the type of the fields

Starting with GDAL 1.6.0, OGR SQL supports changing the type of the columns by using the SQL92 compliant CAST operator according to the following example:

```
SELECT *, CAST(OGR_STYLE AS character(255)) FROM rivers
```

Currently casting to the following target types are supported:

1. boolean (GDAL >= 2.0)
2. character(field_length). By default, field_length=1.
3. float(field_length)
4. numeric(field_length, field_precision)
5. smallint(field_length) : 16 bit signed integer (GDAL >= 2.0)
6. integer(field_length)

7. bigint(field_length), 64 bit integer, extension to SQL92 (GDAL >= 2.0)
8. date(field_length)
9. time(field_length)
10. timestamp(field_length)
11. geometry, geometry(geometry_type), geometry(geometry_type,epsg_code)

Specifying the field_length and/or the field_precision is optional. An explicit value of zero can be used as the width for character() to indicate variable width. Conversion to the 'integer list', 'double list' and 'string list' OGR data types are not supported, which doesn't conform to the SQL92 specification.

While the CAST operator can be applied anywhere in an expression, including in a WHERE clause, the detailed control of output field format is only supported if the CAST operator is the "outer most" operators on a field in the field definition list. In other contexts it is still useful to convert between numeric, string and date data types.

Starting with OGR 1.11, casting a WKT string to a geometry is allowed. geometry_type can be POINT[Z], LINESTRING[Z], POLYGON[Z], MULTIPOINT[Z], MULTILINESTRING[Z], MULTIPOLYGON[Z], GEOMETRYCOLLECTION[Z] or GEOMETRY[Z].

String literals and identifiers quoting

Starting with GDAL 2.0 (see RFC 52 - Strict OGR SQL quoting), strict SQL92 rules are applied regarding string literals and identifiers quoting.

String literals (constants) must be surrounded with single-quote characters. e.g. WHERE a_field = 'a_value'

Identifiers (column names and tables names) can be used unquoted if they don't contain special characters or are not a SQL reserved keyword. Otherwise they must be surrounded with double-quote characters. e.g. WHERE "from" = 5.

WHERE

The argument to the WHERE clause is a logical expression used select records from the source layer. In addition to its use within the WHERE statement, the WHERE clause handling is also used for OGR attribute queries on regular layers via **OGRLayer::SetAttributeFilter()**.

In addition to the arithmetic and other functional operators available in expressions in the field selection clause of the SELECT statement, in the WHERE context logical operators are also available and the evaluated value of the expression should be logical (true or false).

The available logical operators are **=**, **!=**, **<>**, **<**, **>**, **<=**, **>=**, **LIKE** and **ILIKE**, **BETWEEN** and **IN**. Most of the operators are self explanatory, but it is worth noting that **!=** is the same as **<>**, the string equality is case insensitive, but the **<**, **>**, **<=** and **>=** operators *are* case sensitive. Both the LIKE and ILIKE operators are case insensitive.

The value argument to the **LIKE** operator is a pattern against which the value string is matched. In this pattern percent (%) matches any number of characters, and underscore (_) matches any one character. An optional ESCAPE escape_char clause can be added so that the percent or underscore characters can be searched as regular characters, by being preceded with the escape_char.

```
String          Pattern     Matches?
------          -------     --------
Alberta         ALB%        Yes
Alberta         _lberta     Yes
St. Alberta     _lberta     No
St. Alberta     %lberta     Yes
Robarts St.     %Robarts%   Yes
12345           123%45      Yes
123.45          12?45       No
NON 1P0         %NON%       Yes
L4C 5E2         %NON%       No
```

The **IN** takes a list of values as its argument and tests the attribute value for membership in the provided set.

```
Value           Value Set           Matches?
------          -------             --------
321             IN (456,123)        No
'Ontario'       IN ('Ontario','BC') Yes
'Ont'           IN ('Ontario','BC') No
1               IN (0,2,4,6)        No
```

The syntax of the **BETWEEN** operator is "field_name BETWEEN value1 AND value2" and it is equivalent to "field_name >= value1 AND field_name <= value2".

In addition to the above binary operators, there are additional operators for testing if a field is null or not. These are the **IS NULL** and **IS NOT NULL** operators.

Basic field tests can be combined in more complicated predicates using logical operators include **AND**, **OR**, and the unary logical **NOT**. Subexpressions should be bracketed to make precedence clear. Some more complicated predicates are:

```
SELECT * FROM poly WHERE (prop_value >= 100000) AND (prop_value < 200000)
SELECT * FROM poly WHERE NOT (area_code LIKE 'NON%')
SELECT * FROM poly WHERE (prop_value IS NOT NULL) AND (prop_value < 100000)
```

WHERE Limitations

1. Fields must all come from the primary table (the one listed in the FROM clause).
2. All string comparisons are case insensitive except for **<**, **>**, **<=** and **>=**.

ORDER BY

The **ORDER BY** clause is used force the returned features to be reordered into sorted order (ascending or descending) on one of the field values. Ascending (increasing) order is the default if neither the ASC or DESC keyword is provided. For example:

```
SELECT * FROM property WHERE class_code = 7 ORDER BY prop_value DESC
SELECT * FROM property ORDER BY prop_value
SELECT * FROM property ORDER BY prop_value ASC
SELECT DISTINCT zip_code FROM property ORDER BY zip_code
```

Note that ORDER BY clauses cause two passes through the feature set. One to build an in-memory table of field values corresponded with feature ids, and a second pass to fetch the features by feature id in the sorted order. For formats which cannot efficiently randomly read features by feature id this can be a very expensive operation.

Sorting of string field values is case sensitive, not case insensitive like in most other parts of OGR SQL.

LIMIT and OFFSET

Starting with GDAL 2.2, the **LIMIT** clause can be used to limit the number of features returned. For example

```
SELECT * FROM poly LIMIT 5
```

The **OFFSET** clause can be used to skip the first features of the result set. The value after OFFSET is the number of features skipped. For example, to skip the first 3 features from the result set:

```
SELECT * FROM poly OFFSET 3
```

Both clauses can be combined:

```
SELECT * FROM poly LIMIT 5 OFFSET 3
```

JOINs

OGR SQL supports a limited form of one to one JOIN. This allows records from a secondary table to be looked up based on a shared key between it and the primary table being queried. For instance, a table of city locations might include a *nation_id* column that can be used as a reference into a secondary *nation* table to fetch a nation name. A joined query might look like:

```
SELECT city.*, nation.name FROM city
     LEFT JOIN nation ON city.nation_id = nation.id
```

This query would result in a table with all the fields from the city table, and an additional "nation.name" field with the nation name pulled from the nation table by looking for the record in the nation table that has the "id" field with the same value as the city.nation_id field.

Joins introduce a number of additional issues. One is the concept of table qualifiers on field names. For instance, referring to city.nation_id instead of just nation_id to indicate the nation_id field from the city layer. The table name qualifiers may only be used in the field list, and within the **ON** clause of the join.

Wildcards are also somewhat more involved. All fields from the primary table (*city* in this case) and the secondary table (*nation* in this case) may be selected using the usual ***** wildcard. But the fields of just one of the primary or secondary table may be selected by prefixing the asterix with the table name.

The field names in the resulting query layer will be qualified by the table name, if the table name is given as a qualifier in the field list. In addition field names will be qualified with a table name if they would conflict with earlier fields. For instance, the following select would result might result in a results set with a *name, nation_id, nation.nation_id* and *nation.name* field if the city and nation tables both have the *nation_id* and *name* fieldnames.

```
SELECT * FROM city LEFT JOIN nation ON city.nation_id = nation.nation_id
```

On the other hand if the nation table had a *continent_id* field, but the city table did not, then that field would not need to be qualified in the result set. However, if the selected instead looked like the following statement, all result fields would be qualified by the table name.

```
SELECT city.*, nation.* FROM city
    LEFT JOIN nation ON city.nation_id = nation.nation_id
```

In the above examples, the *nation* table was found in the same datasource as the *city* table. However, the OGR join support includes the ability to join against a table in a different data source, potentially of a different format. This is indicated by qualifying the secondary table name with a datasource name. In this case the secondary datasource is opened using normal OGR semantics and utilized to access the secondary table until the query result is no longer needed.

```
SELECT * FROM city
  LEFT JOIN '/usr2/data/nation.dbf'.nation ON city.nation_id = nation.nation_id
```

While not necessarily very useful, it is also possible to introduce table aliases to simplify some SELECT statements. This can also be useful to disambiguate situations where tables of the same name are being used from different data sources. For instance, if the actual tables names were messy we might want to do something like:

```
SELECT c.name, n.name FROM project_615_city c
  LEFT JOIN '/usr2/data/project_615_nation.dbf'.project_615_nation n
            ON c.nation_id = n.nation_id
```

It is possible to do multiple joins in a single query.

```
SELECT city.name, prov.name, nation.name FROM city
  LEFT JOIN province ON city.prov_id = province.id
  LEFT JOIN nation ON city.nation_id = nation.id
```

Before GDAL 2.0, the expression after ON should necessarily be of the form "{primary_table}.{field_name} = {secondary_table}.{field_name}", and in that order. Starting with GDAL 2.0, it is possible to use a more complex boolean expression, involving multiple comparison operators, but with the restrictions mentioned in the below "JOIN limitations" section. In particular, in case of multiple joins (3 tables or more) the fields compared in a JOIN must belong to the primary table (the one after FROM) and the table of the active JOIN.

JOIN Limitations

1. Joins can be very expensive operations if the secondary table is not indexed on the key field being used.
2. Joined fields may not be used in WHERE clauses, or ORDER BY clauses at this time. The join is essentially evaluated after all primary table subsetting is complete, and after the ORDER BY pass.
3. Joined fields may not be used as keys in later joins. So you could not use the province id in a city to lookup the province record, and then use a nation id from the province id to lookup the nation record. This is a sensible thing to want and could be implemented, but is not currently supported.
4. Datasource names for joined tables are evaluated relative to the current processes working directory, not the path to the primary datasource.
5. These are not true LEFT or RIGHT joins in the RDBMS sense. Whether or not a secondary record exists for the join key or not, one and only one copy of the primary record is returned in the result set. If a secondary record cannot be found, the secondary derived fields will be NULL. If more than one matching secondary field is found only the first will be used.

UNION ALL

(OGR >= 1.10.0)

The SQL engine can deal with several SELECT combined with UNION ALL. The effect of UNION ALL is to concatenate the rows returned by the right SELECT statement to the rows returned by the left SELECT statement.

```
[(] SELECT field_list FROM first_layer [WHERE where_expr] [)]
UNION ALL [(] SELECT field_list FROM second_layer [WHERE where_expr] [)]
[UNION ALL [(] SELECT field_list FROM third_layer [WHERE where_expr] [)]]*
```

UNION ALL restrictions

The processing of UNION ALL in OGR differs from the SQL standard, in which it accepts that the columns from the various SELECT are not identical. In that case, it will return a super-set of all the fields from each SELECT statement.

There is also a restriction : ORDER BY can only be specified for each SELECT, and not at the level of the result of the union.

SPECIAL FIELDS

The OGR SQL query processor treats some of the attributes of the features as built-in special fields can be used in the SQL statements likewise the other fields. These fields can be placed in the select list, the WHERE clause and the ORDER BY clause respectively. The special field will not be included in the result by default but it may be explicitly included by adding it to the select list. When accessing the field values the special fields will take precedence over the other fields with the same names in the data source.

FID

Normally the feature id is a special property of a feature and not treated as an attribute of the feature. In some cases it is convenient to be able to utilize the feature id in queries and result sets as a regular field. To do so use the name **FID**. The field wildcard expansions will not include the feature id, but it may be explicitly included using a syntax like:

```
SELECT FID, * FROM nation
```

OGR_GEOMETRY

Some of the data sources (like MapInfo tab) can handle geometries of different types within the same layer. The **OGR_GEOMETRY** special field represents the geometry type returned by **OGRGeometry::getGeometryName()** and can be used to distinguish the various types. By using this field one can select particular types of the geometries like:

```
SELECT * FROM nation WHERE OGR_GEOMETRY='POINT' OR OGR_GEOMETRY='POLYGON'
```

OGR_GEOM_WKT

The Well Known Text representation of the geometry can also be used as a special field. To select the WKT of the geometry **OGR_GEOM_WKT** might be included in the select list, like:

```
SELECT OGR_GEOM_WKT, * FROM nation
```

Using the **OGR_GEOM_WKT** and the **LIKE** operator in the WHERE clause we can get similar effect as using OGR_GEOMETRY:

```
SELECT OGR_GEOM_WKT, * FROM nation WHERE OGR_GEOM_WKT
   LIKE 'POINT%' OR OGR_GEOM_WKT LIKE 'POLYGON%'
```

OGR_GEOM_AREA

(Since GDAL 1.7.0)

The **OGR_GEOM_AREA** special field returns the area of the feature's geometry computed by the **OGRSurface::get_Area()** method. For **OGRGeometryCollection** and **OGRMultiPolygon** the value is the sum of the areas of its members. For non-surface geometries the returned area is 0.0.

For example, to select only polygon features larger than a given area:

```
SELECT * FROM nation WHERE OGR_GEOM_AREA > 10000000
```

OGR_STYLE

The **OGR_STYLE** special field represents the style string of the feature returned by **OGRFeature::GetStyleString()**. By using this field and the **LIKE** operator the result of the query can be filtered by the style. For example we can select the annotation features as:

```
SELECT * FROM nation WHERE OGR_STYLE LIKE 'LABEL%'
```

CREATE INDEX

Some OGR SQL drivers support creating of attribute indexes. Currently this includes the Shapefile driver. An index accelerates very simple attribute queries of the form *fieldname = value*, which is what is used by the **JOIN** capability. To create an attribute index on the nation_id field of the nation table a command like this would be used:

```
CREATE INDEX ON nation USING nation_id
```

Index Limitations

1. Indexes are not maintained dynamically when new features are added to or removed from a layer.
2. Very long strings (longer than 256 characters?) cannot currently be indexed.
3. To recreate an index it is necessary to drop all indexes on a layer and then recreate all the indexes.
4. Indexes are not used in any complex queries. Currently the only query the will accelerate is a simple "field = value" query.

DROP INDEX

The OGR SQL DROP INDEX command can be used to drop all indexes on a particular table, or just the index for a particular column.

```
DROP INDEX ON nation USING nation_id
DROP INDEX ON nation
```

ALTER TABLE

(OGR >= 1.9.0)

The following OGR SQL ALTER TABLE commands can be used.

1. "ALTER TABLE tablename ADD [COLUMN] columnname columntype" to add a new field. Supported if the layer declares the OLCCreateField capability.
2. "ALTER TABLE tablename RENAME [COLUMN] oldcolumnname TO newcolumnname" to rename an existing field. Supported if the layer declares the OLCAlterFieldDefn capability.
3. "ALTER TABLE tablename ALTER [COLUMN] columnname TYPE columntype" to change the type of an existing field. Supported if the layer declares the OLCAlterFieldDefn capability.
4. "ALTER TABLE tablename DROP [COLUMN] columnname" to delete an existing field. Supported if the layer declares the OLCDeleteField capability.

The columntype value follows the syntax of the types supported by the CAST operator described above.

```
ALTER TABLE nation ADD COLUMN myfield integer
ALTER TABLE nation RENAME COLUMN myfield TO myfield2
ALTER TABLE nation ALTER COLUMN myfield2 TYPE character(15)
ALTER TABLE nation DROP COLUMN myfield2
```

DROP TABLE

(OGR >= 1.9.0)

The OGR SQL DROP TABLE command can be used to delete a table. This is only supported on datasources that declare the ODsCDeleteLayer capability.

```
DROP TABLE nation
```

ExecuteSQL()

SQL is executed against an **GDALDataset**, not against a specific layer. The call looks like this:

```
OGRLayer * GDALDataset::ExecuteSQL( const char *pszSQLCommand,
                                    OGRGeometry *poSpatialFilter,
                                    const char *pszDialect );
```

The pszDialect argument is in theory intended to allow for support of different command languages against a provider, but for now applications should always pass an empty (not NULL) string to get the default dialect.

The poSpatialFilter argument is a geometry used to select a bounding rectangle for features to be returned in a manner similar to the **OGRLayer::SetSpatialFilter()** method. It may be NULL for no special spatial restriction.

The result of an ExecuteSQL() call is usually a temporary **OGRLayer** representing the results set from the statement. This is the case for a SELECT statement for instance. The returned temporary layer should be released with GDALDataset::ReleaseResultsSet() method when no longer needed. Failure to release it before the datasource is destroyed may result in a crash.

Non-OGR SQL

All OGR drivers for database systems: MySQL, PostgreSQL and PostGIS (PG), Oracle (OCI), SQLite, ODBC, ESRI Personal Geodatabase (PGeo) and MS SQL Spatial (MSSQLSpatial), override the **GDALDataset::ExecuteSQL()** function with dedicated implementation and, by default, pass the SQL statements directly to the underlying RDBMS. In these cases the SQL syntax varies in some particulars from OGR SQL. Also, anything possible in SQL can then be accomplished for these particular databases. Only the result of SQL WHERE statements will be returned as layers.
