

Supporting Table Aliases and Qualified Column Names in DataFusion

Related JIRA: [ARROW-10761](#), [ARROW-10732](#), [ARROW-10813](#)

Objective

It should be possible to join two relations that have one or more columns with the same name and then use qualified names to disambiguate when referencing the output columns from the join.

Apache Spark supports table aliases and compound identifiers with both the DataFrame and SQL APIs. Let's use the following two tables for the following examples:

```
val df1 = Seq(("aa", "11"), ("bb", "22"), ("cc", "33")).toDF("id", "name")
val df2 = Seq(("aa", "99"), ("bb", "88"), ("cc", "77")).toDF("id", "name")
```

Here as an example from Apache Spark using the DataFrame API:

```
val df3 = df1.as("t1").join(df2.as("t2"), col("t1.id").equalTo(col("t2.id")))
val df4 = df3.select("t1.id", "t2.name")
```

Here is the same example but using the SQL API.

```
df1.createOrReplaceTempView("t1")
df2.createOrReplaceTempView("t2")
val df3 = spark.sql("SELECT t1.id, t2.name FROM t1 JOIN t2 ON t1.id = t2.id")
```

Proposal

I originally thought that we could isolate changes to the SQL planner in DataFusion, but this is not possible because the SQL planner delegates to the LogicalPlanBuilder (which seems like a good design because it ensures consistency between the DataFrame and SQL APIs). Also, I now think that we do want to be able to have the same behavior available from the DataFrame API.

I now think we should make the following changes:

- Introduce new SQLSchema / SQLField structs that wrap the core Arrow Schema / Field structs but add support for qualified field names
- Update LogicalPlan to use SQLSchema instead of Schema

- LogicalPlanBuilder/SQLSchema should contain logic to find fields in schemas based on qualified or unqualified column names, and in the unqualified case it should ensure that the reference is not ambiguous
- Update Expr::Column to use Vec<String> rather than String for column names
- When the query is executed, qualifiers should be stripped from the final schema, and there should be a check to ensure that the output column names are unique