Spark SQL: Relational Data Processing in Spark

I. Summary

evaluates operations lazily so that it can perform relational optimizations.

DataFrame API

extensible query optimizer called Catalyst.

Spark SQL runs as a library on top of Spark.

A DataFrame is equivalent to a table in a relational database, lazy

DSL domain-specific language AST abstract syntax tree

ORM object-relational mapping

Spark SQL can cache hot data in memory using columnar storage.

UDF User-defined functions

Trees can be manipulated using rules, which are functions from a tree to another tree.

analyzing a logical plan to resolve references, logical plan optimization, physical planning, code generation to compile parts of the query to Java bytecode.

An attribute is called unresolved if we do not know its type or have not matched it to a