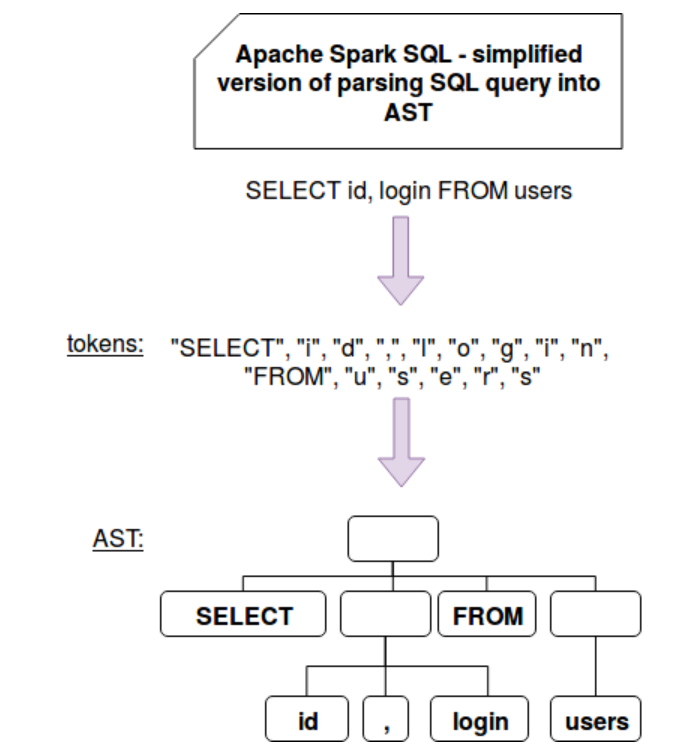
**SQL text to Logical Plan:**

**What happens when we execute a sql query in Spark?**

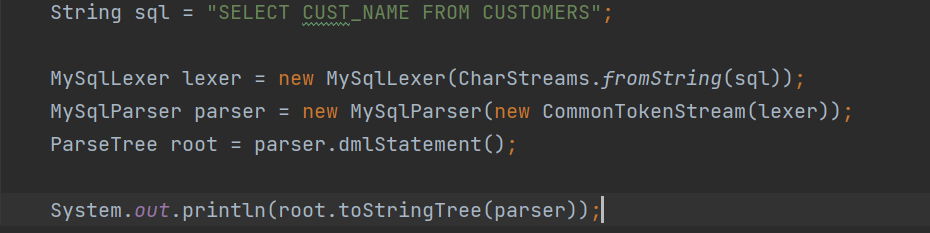
When we write a Spark program to execute a sql query “SELECT CUST\_NAME FROM CUSTOMERS”( for an example), Spark first does lexical analysis in order to check syntax and then creates token from that query text “SELECT CUST\_NAME FROM *CUSTOMERS”. This lexical Analysis is done by* lexer. This Process also called as “Tokenization”. This Token will includes Operators “SELECT”, “FROM” and “WHERE” etc.,

And then Parser (SqlBaseParser) creates Abstract Syntax Tree(AST), from the tokens, that will be later interpreted by Spark to build the logical plan using AstBuilder class.



Spark uses library called ANTLR(Another Tool for Language Recognition)

Sample Java Program for Lexer:



AST Output of above Java Code:

*(dmlStatement (selectStatement (querySpecification SELECT (selectElements (selectElement (fullColumnName (uid (simpleId CUST\_NAME))))) (fromClause FROM (tableSources (tableSource (tableSourceItem (tableName (fullId (uid (simpleId CUSTOMERS)))))))))))*

Same as above Java Program, Spark uses AbstractSqlParser for lexer and Parser in order to generate AST,

**protected def parse[T](command: String)(toResult: SqlBaseParser => T): T = {**

**logDebug(s"Parsing command: $command")**

**val lexer = new SqlBaseLexer(new UpperCaseCharStream(CharStreams.fromString(command)))**

**..**

**..**

**val tokenStream = new CommonTokenStream(lexer)**

**val parser = new SqlBaseParser(tokenStream)**

**..**

**..**

**}**

The syntax definition is included in a .g4 file which is also called a grammar file. In other terms, it's nothing more than a set of lexer and parser rules.

(grammar: org\apache\spark\sql\catalyst\parser\SqlBase.g4)

SqlBaseLexer is used for Lexer. Creates Tokens

SqlBaseParser is used for Parser. Builds AST tree and Logical Plan

CommonTokenStream for tokens. Used by Lexer.

Ref:

[1] <https://www.waitingforcode.com/apache-spark-sql/writing-custom-optimization-apache-spark-sql-parser/read>

[2] https://docs.ksqldb.io/en/latest/operate-and-deploy/how-it-works/