

Final Course Project Report

SPARK API Converter

Md Rayhanur Rahman*

Ph.D. Student
Dept. of CSC
NC State University
Raleigh, NC, USA
mrahman@ncsu.edu

Mohammad Maruful Haque†

Ph.D. Student
Dept. of CSC
NC State University
Raleigh, NC, USA
mhaque3@ncsu.edu

Abstract

This document contains the final report of the Spark API Converter project which is a course project requirements of CSC512 - Compiler Construction course. The documents highlights the problem definition, motivation, background, related work, methodology, workload distribution, results, limitations and future work direction.

1 The Problem

In this project, we need to build a source code converter which can transform these following things along with some additional tasks.

1. **RQ1:** Getting introduced with Apache Spark, setting it up in a docker container and writing some piece of code in Spark RDD, Dataset and Dataframe API. In the project requirement document, it is referred as Part 0
2. **RQ2:** Transforming Spark RDD API to Spark Dataset API. In the project requirement document, it is referred as Part 1
3. **RQ3:** Transforming Spark RDD API to Spark Dataframe API. In the project requirement document, it is referred as Part 2
4. **RQ4:** Implementing some Spark SQL API calls to Scala functions and providing thoughts on how to implement a generic transformer that can transform Scala functions to corresponding SQL code. In the project requirement document, it is referred as Part 3

2 Background and Motivation

Spark is very popular these days for distributed computing tasks such as data analysis, graph modeling and machine learning etc. There are three APIs in Spark at present named RDD, Dataset and Dataframe. The three different APIs provide different expressive power to achieve the same thing although the performance gain can be different. Moreover, Spark introduced these three APIs in different times so that

there are many existing implementations of Spark based systems that needs to be rewritten to gain better performance. So, an API transformer can be handy to migrate Spark APIs to achieve better performance. Moreover, by doing this, a great understanding will be built on compiler techniques such as scanning and parsing, handling CFG grammar and intermediate representation.

2.1 Motivation

The problem we are about to work upon largely falls into the domain of source to source compiler or transcompiler. It is a type of compiler where the source code of a program written in one programming language is taken as input and the equivalent source code in another programming language is produced as an output. It is different from traditional compiler in a sense that source-to-source compilers converts one programming language to another where both of those belongs to the same abstraction level whereas a conventional compiler turns a high level language to a low level one. These types of compiler is also useful in case of

- transforming legacy code base to a modern counterpart such as structured to OOP design [10]
- transforming to newer set of APIs. For example, .net framework and .net core framework [4]
- maintaining issues with backward compatibility, for example, running iOS apps to macOS
- transforming code in more modular version
- porting a code to different platforms [3]
- reducing technical debt of code base by applying automatic refactoring
- developing new language on top of existing and established language such as *typescript* running on top of *javascript* [5]

In our context, it is already known that, in terms of performance, *dataframe* > *dataset* > *RDD*. Hence, a legacy distributed computing focused application written in *Spark RDD* API can be transformed to *Dataset* or *Dataframe* API to extract the best performance from the computing devices. If we can build an automatic transcompiler that can perform aforementioned duty, then it could save a considerable amount of human effort and brings out better performance at the same time.

*unity id: 200255928

†unity id: 200262103

3 Related Work

Source-to-Source transformation has been treated with importance among researchers for its contribution to improve the program performance [7]. There has been study regarding generation of dependable software and application using the technique of source-to-source compiler [1] [8]. Tools like Cetus [2] has been introduced supporting source-to-source transformations on C programs providing parallelization passes with multicore compiler optimizations. Researchers have also presented scholarly article with a compiler framework for automatic source-to-source translation of standard OpenMP applications into CUDA-based GPGPU applications [6]. There they have identified several key transformation techniques, which enable efficient GPU global memory access, to achieve high performance. Article has also been published [9] that translates C code with OpenACC directives to C code with the CUDA API adopting source-to-source approach using the Omni compiler infrastructure for source code analysis and translations.

4 Methodology

In this section, the step by step process for each of the task is described below:

4.1 RQ1

We have installed a Docker container with spark shell in it. This spark-shell is built upon Scala, so Spark's APIs for Scala are integrated with this container. We used this environment to get familiar with Spark's RDD, Dataset and Dataframe API.

4.1.1 Program implementation

We have also implemented 6 different functions with each one of Spark's APIs with a subtotal of 18 programs. Following is a sample of 3 implementations for each of RDD, Dataset and Dataframe.

```
sc.range(1,100)
  .filter(i => (i % 5 == 0))
  .reduce((a:Long, b:Long) => a+b)

spark.range(1,100).as[Long]
  .filter(i => (i % 5 == 0))
  .select(reduceAggregator(
    (a:Long, b:Long) => a+b))
  .collect()

spark.range(1,100).selectExpr("id as _1")
  .selectExpr("if(_1%5==0,_1,_1*0) as _1")
  .selectExpr("sum(_1)")
  .as[Long].first
```

| | |
|------------------|----------------------|
| Operating System | Arch Linux |
| Kernel Version | 4.19 |
| Python Version | 3.7 |
| Processor | Core i7 8550U 1.8GHz |
| Memory | 16 GB 2400 MHz |
| Disk | 240GB SSD |

Table 1. Execution Environment for RQ2

4.2 RQ2

4.2.1 Execution Environment

Here, we implemented a python program that can handle the requirements stated in RQ2. Our machine was a typical laptop computer. Here is the description of the execution environment.

4.2.2 Libraries

We used raw python codes to do the task. No additional library modules were used to achieve the goal for this task.

4.2.3 The Goal

For this following input:

```
sc.range(1,10000000)
  .map(i => (i % 11, 1))
  .reduceByKey((a: Int, b: Int) => a+b)
  .collect()
```

We need to produce the following output:

```
spark.range(1,10000000)
  .map(i => (i % 11, 1))
  .groupByKey(_._1)
  .agg(reduceByKeyAggregator((a: Int, b: Int)
    => a+b))
  .collect()
```

4.2.4 Algorithm

Here is the step by step process on how we achieved the requirements of RQ2.

| | |
|---------------------|--|
| sc | Spark |
| range | range |
| textFile | read.textFile |
| map | map |
| filter | filter |
| reduce(<func>) | select(reduceAggregator(<func>)).collect() |
| reduceByKey(<func>) | groupByKey(_._1).agg(reduceByKeyAggregator(<func>)) |
| sortBy(<func>) | map(row=>(<func>(row), row)).orderBy("_1").map(_._2) |
| collect | collect |

Figure 1. Transformation Rules

Listing 1. token definition

<letter> --> a | b | ... | y | z

| A | B | ... | Z |
 <digit> --> 0 | 1 | ... | 9
 <number> --> <digit>+
 <identifier> --> <letter> (<letter>
 | <digit>)*
 <string> --> any string between (and
 including) the closest pair of double
 quotation marks.
 <char> --> a character between (and
 including) a pair of single quotation
 marks.
 <symbol> --> any non-space character
 that is not a part of other tokens

1. First we took the whole input as a simple python string object.
2. We filtered all the new line, carriage returns and tabs from that string
3. We wrote regular expression checker using python's default *re* module for the following tokens highlighted in list 1
4. After that, we read the whole string one character at a time and tokenized the whole string
5. we replaced the tokens according to the transformation rules mentioned above
6. just replacing the tokens does not actually do the whole job. We needed to match the opening brackets and closing brackets associated with each of the tokens and then we replaced the tokens and appended corresponding tokens after the right parenthesis according to the transformation rules

4.2.5 Sample Test Cases

A sample input-output is provided here.

Input (RDD)

```
sc.textFile("/home/sample.txt")
  .map(line => line.split(" ").size)
  .reduce((a: Int, b: Int) =>
    if (a > b) a else b)
```

Output (Dataset)

```
spark.read.textFile("/home/sample.txt")
  .map(line => line.split(" ").size)
  .select(reduceAggregator((a: Int, b: Int)
    => if (a > b) a else b)).collect()
```

4.2.6 Discussions

Here is some of the observation we came through while implementing the RQ2.

- Our transformed would not work if there will be any identifier having names that is present in the transformation table. For example, this following statement will produce an error.

```
map(x => sc)
```

Because *sc* is already in the transformation table.

- This phenomenon happens because the transformed would not know if the *sc* is an identifier or an object. This aforementioned problem could have been solved if we could have constructed the whole grammar for the Spark RDD API. But developing a full fledged grammar was an infeasible task, hence we did not implemented it. If the grammar was there, then it was fully possible to first parsing the tokens, building an abstract syntax tree and replace the nodes with corresponding nodes according to the transformation rules.

4.3 RQ3

4.3.1 Execution Environment

The environment details are as same as Table 1.

4.3.2 Libraries

Initially we planned to build the whole transformer, (i.e. scanner, parser, tree) from scratch. Later, we found out that there are some existing excellent framework for doing these types of tasks with *ANTLR*, *YACC*, *JavaCC*, *Scala Parser Combinators* etc. However, finally we choose a framework named *Lark* which is written in pure Python to generate scanners and parsers for our goal. Lark can:

- Parse all context-free grammars, and handle all ambiguity
- Build a parse-tree automatically, no construction code required
- Outperform all other Python libraries when using LALR(1) (Yes, including PLY) Run on every Python interpreter (it's pure-python) Generate a stand-alone parser (for LALR(1) grammars)

We didn't choose the De Facto *ANTLR* as it is heavy and initial learning curving is a bit steep despite being a powerful framework for these type of tasks.

4.3.3 The Goal

For the following input,

```
sc.range(10,100)
  .map(i=>{val j=i%3;
    (i, if(j==0)i*10 else i*2)})
  .map(r=>r._1+r._2)
  .collect()
```

We have to produce the following output,

```
spark.range(10,100).selectExpr("id as _1")
  .selectExpr("_1 as _1",
```

```
" if (_1%3==0,_1*10,_1*2) as _2")
.selectExpr("_1+_2 as _1")
.collect()
```

4.3.4 Algorithm

The step by step process to achieve the goal stated in RQ3 is mentioned below.

Listing 2. UDF Grammar

```
<Program> ::= sc.range(<number>,<number>)
<MapOps>.collect()
<MapOps> ::= | <MapOps>.map(<UDF>)
<UDF> ::= <identifier> => <Expression>
<Expression> ::= {<ComplexExpr>}
| <SimpleExpr>
<SimpleExpr> ::= <PureExpr>
| (<TupleExpr>)
<TupleExpr> ::= <PureExpr>,<PureExpr>
| <TupleExpr>,<PureExpr>
<ComplexExpr> ::= <SimpleExpr>
| <AssignExprs>;<SimpleExpr>
<AssignExprs> ::= <AssignExpr>
| <AssignExprs>;<AssignExpr>
<AssignExpr> ::= val <identifier> =
<PureExpr>
<PureExpr> ::= <identifier>
| <identifier>.<identifier>
| (<PureExpr>)
| <PureExpr> <Op> <PureExpr>
| if (<CompExpr>) <PureExpr> else
<PureExpr>
<CompExpr> ::= <PureExpr> <Comp>
<PureExpr>
<Op> ::= + | - | * | %
<Comp> ::= == | < | > | != | >= | <=
```

Listing 3. lark grammar

```
// rdd grammar
start : SC actions*
actions : DOT RANGE LP rangeparams* RP
| DOT TEXTFILE LP URI RP
| DOT MAP LP func RP
| DOT FILTER LP func RP
| DOT REDUCE LP func RP
| DOT REDUCEBYKEY LP func RP
| DOT SORTBY LP func RP
| DOT COLLECT LP RP

rangeparams : NUMBER (COMMA NUMBER)?
```

```
func : ID ARROW expression
expression : simpleexpression
| LB complexexpression RB
simpleexpression : pureexpression
| LP tupleexpression RP
tupleexpression : pureexpression COMMA
pureexpression
| tupleexpression COMMA
pureexpression
complexexpression : simpleexpression
| assignmentexpressions
SEMICOLON
simpleexpression
assignmentexpressions :
assignmentexpression
| assignmentexpressions SEMICOLON
assignmentexpression :
assignmentexpression :
VAL ID EQUAL pureexpression
pureexpression : NUMBER
| ID
| ID DOT ID
| LP pureexpression RP
| pureexpression OP
pureexpression
| IF LP
comparisonexpression
RP pureexpression
ELSE pureexpression
comparisonexpression : pureexpression
COMP pureexpression
```

```
SC : "sc"
VAL : "val"
IF : "if"
ELSE : "else"
EQUAL : "="
DOT : /[.]/
OP : /[ ]*["+\-""%"]/[ ]*/
COMP : ([ ]*[>=]=)[ ]*/
| /[ ]*(>)[ ]*/ | /[ ]*(<)[ ]*/
| /[ ]*(!)[ ]*/ | /[ ]*(>)[ ]*/
| /[ ]*(<)[ ]*/ | /[ ]*(<)[ ]*/
COMMA : /[,]/
SEMICOLON : /[;]/
ARROW : /[ ]*=[ ]*>[ ]*/
LB : /[ ]*{[ ]*/
```

```

RB : /[ ]*{}[ ]*/
LP : /[ ]*"("[ ]*[/
RP : /[ ]*"")[/ ]*/
RANGE : "range"
TEXTFILE : "textFile"
MAP : "map"
FILTER : "filter"
REDUCE : "reduce"
REDUCEBYKEY : "reduceByKey"
SORTBY : "sortBy"
COLLECT : "collect"

URI : ( /[ "\ " ]/ | /[ " ' " ]/ )
/[ \\ a-zA-Z . : \ / ] * /
( /[ "\ " ]/ | /[ " ' " ]/ )
//FUNC : "<func>"

%import common.CNAME -> ID
%import common.SIGNED_NUMBER
-> NUMBER
%import common.WS
%ignore WS

```


| | |
|------------|---|
| sc | spark |
| range(m,n) | range(m,n).selectExpr("id as _1") |
| map(UDF) | selectExpr(SQL)  |
| collect | collect |

Figure 2. Transformation Rules

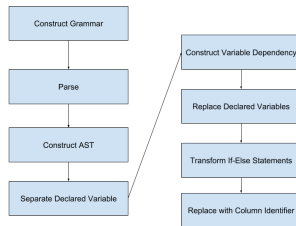


Figure 3. Flowchart of RQ3

1. First, we need to have a grammar upon which the parser and abstract syntax tree will be built upon. The grammar has been given in the list 2
2. As we used *lark-parser* for the task of parsing and generating abstract syntax tree, we needed to transform the grammar to feed into the *lark-parser* convention. That grammar is given in list 3
3. we developed an abstract syntax tree on top of the *lark-parser*.

4. we extracted the tokens which were needed to be replaced according to the transformation rules given in figure 2
5. the most challenging part is obviously transforming the UDF to SQL. There are actually two source of difficulties in transforming UDF to SQL. First one is to figure out the dependencies of the variables of *Assign-Exprs*. The second one is to convert the *if-else* portion of the UDF to corresponding SQL syntax.
6. For the first challenge, we extracted the UDF portion from the input first. Then we separated the variable declaration part from the rest of the expression. Then we determined the dependency of the variables and replaced all the variables with appropriate mathematical expression. At the end, all the declared variables were replaced.
7. For the second challenge, we parsed the *if-else* and nested *if-else* statements using a *stack* based data structures. Then we replaced all the variables of conditions with the root variable that came from the input of the lambda expression.
8. Finally, we replaced the input variables with Spark SQL column identifier such as *_1, _2* etc.

The whole step by step process is also represented as flowchart in figure 3.

4.3.5 Sample Test Cases

A sample input-output of RDD to Dataframe is provided here.

Input (RDD)

```

sc.range(1,10)
  .map(x => (x+2, if(x+3 > 6) x else 1))
  .collect()

```

Output (Dataframe)

```

spark.range(1,10)
  .selectExpr("id as _1")
  .selectExpr("_1 + 2 as _1",
    "if(_1 + 3 > 6, _1, 1) as _2")
  .collect()

```

4.3.6 Discussion

Here is some observation we came through while implementing RQ3.

- Our implementation only works for the simple UDF functions as provided in the grammar described in the project requirements
- Our implementation contains parenthesis to preserve the meaning of mathematical formulas. In some cases, those brackets were redundant though.
- Handling the nested *if-else* was particularly challenging.

- We assume that, there might be a more elegant solution to achieve the goal of RQ3. For example, the whole thing could have done recursively using abstract syntax tree and intermediate code representation. For example, a directed acyclic graph could have been drawn to identify the data flow and variable dependencies.
- Moreover, the nested *if-else* statements could have automatically constructed with the abstract syntax tree. After that, a recursive transformer could have converted those expression to corresponding SQL statements.

4.4 RQ4

We have selected 10 Spark SQL APIs for Scala implementation from the provided list and the APIs are `concat`, `hex(str)`, `unhex(str)`, `repeat`, `rtrim`, `ltrim`, `rpadd`, `lpadd`, `translate` and `reverse`. Following is the sample of Scala code for Spark SQL API `lpadd`.

```
def lpadd(str: String, len: Int,
          pad: String) = {
    var strLength = str.length()
    var diff = len - strLength
    var ret = str
    if (diff < 0){
        ret = str.substring(0, len)
    }
    if(diff > 0 ){
        while(ret.length() < len) {
            ret = pad + ret
        }
        var ret_length = ret.length()
        if(ret_length > len){
            ret = ret.substring
                (ret_length - len, ret_length)
        }
    }
    ret
}
```

4.4.1 How a Generic Scala to SQL Transformer can be Built

Although, in this course project, we have implemented a simple transformer that can basically transform simple UDF in lambda queries to Spark SQL syntax. But, that can be generalized to whole bunch of full fledged scala lambda queries to SQL. In order to do that, we can build an expression tree which will contain each lambda expressions *left*, *operator*, *right* and the *node type* information. Expressions can be of these following types:

- **unary:** An operation with a single operand such as negation

- **binary:** An operation with two operands such as addition or `||`.
- **parameter:** An input to a lambda function
- **member:** Accessing a property, field, or method of an object or a variable.
- **constant:** A node that is a constant value

After constructing the tree, we can recursively traverse each nodes and generate the equivalent SQL queries for each nodes. For example, if a node contains *startsWith* method, we can generate the SQL query contain *LIKE* operator. However, it might be the case that not all Scala operations are automatically transformable to the corresponding SQL queries, i.e., custom functions.

5 How to setup Testing Environment

The transformer program is standalone and self packaged. Hence, there is no need for configuration or settings file. It is designed in such a way so that the tester can test it by just running the scripts and providing appropriate input files. Custom test cases can be written as well to test some sample inputs and outputs. The code samples as input and produced output code samples must preserve the same meaning and produce the same result according to the transformation table. That is the key testing criteria. Our test programs for Part 0 and Part 3 are written in Scala. For the part 1 and part 2, the test programs are written in Python. Here are the requirements to run the test programs.

- Linux or Mac OSX
- Docker container where Apache Spark has been setup
- Python3
- *lark-parser* python package

6 Workload Distribution

After completing the whole task, here is the workload distribution for our team.

- Part 0: Done by Maruful
- Part 1: Done by Rayhanur
- Part 2: Done by Rayhanur and Maruful
- Part 3: Done by Maruful and Rayhanur

References

- [1] A. Benso, S. Chiusano, P. Prinetto, and L. Tagliaferri. 2000. A C/C++ source-to-source compiler for dependable applications. In *Proceeding International Conference on Dependable Systems and Networks. DSN 2000*. 71–78. <https://doi.org/10.1109/ICDSN.2000.857517>
- [2] C. Dave, H. Bae, S. Min, S. Lee, R. Eigenmann, and S. Midkiff. 2009. Cetus: A Source-to-Source Compiler Infrastructure for Multicores. *Computer* 42, 12 (Dec 2009), 36–42. <https://doi.org/10.1109/MC.2009.385>
- [3] Facebook Inc. 2018. ReactNative. <https://facebook.github.io/react-native/>. [Online; last accessed 9-Dec-2008].
- [4] Microsoft Inc. 2018. Port your code from .NET Framework to .NET Core. <https://docs.microsoft.com/en-us/dotnet/core/porting/>. [Online; last accessed 9-Dec-2008].

- [5] Microsoft Inc. 2018. Typescript. <https://www.typescriptlang.org/>. [Online; last accessed 9-Dec-2008].
- [6] Seyong Lee, Seung-Jai Min, and Rudolf Eigenmann. 2009. OpenMP to GPGPU: A Compiler Framework for Automatic Translation and Optimization. In *Proceedings of the 14th ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming (PPoPP '09)*. ACM, New York, NY, USA, 101–110. <https://doi.org/10.1145/1504176.1504194>
- [7] David B. Loveman. 1977. Program Improvement by Source-to-Source Transformation. *J. ACM* 24, 1 (Jan. 1977), 121–145. <https://doi.org/10.1145/321992.322000>
- [8] M. Rebaudengo, M. S. Reorda, M. Violante, and M. Torchiano. 2001. A source-to-source compiler for generating dependable software. In *Proceedings First IEEE International Workshop on Source Code Analysis and Manipulation*. 33–42. <https://doi.org/10.1109/SCAM.2001.972664>
- [9] Akihiro Tabuchi, Masahiro Nakao, and Mitsuhsa Sato. 2014. A Source-to-Source OpenACC Compiler for CUDA. In *Euro-Par 2013: Parallel Processing Workshops*, Dieter an Mey, Michael Alexander, Paolo Bientinesi, Mario Cannataro, Carsten Clauss, Alexandru Costan, Gabor Kecskemeti, Christine Morin, Laura Ricci, Julio Sahuquillo, Martin Schulz, Vittorio Scarano, Stephen L. Scott, and Josef Weidendorfer (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 178–187.
- [10] Ying Zou and Kostas Kontogiannis. 2001. A framework for migrating procedural code to object-oriented platforms. In *apsec*. IEEE, 390.