Course Project Writeup - 0*

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 report contains the initial concept notes from the SCALA API Converter programming project describing the problem definition, background, related work, methodology, timeline, testing criteria and finally workload distribution.

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. 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
2. Spark RDD API to Spark Dataset API
3. Spark RDD API to Spark Dataframe API
4. Implementing some Spark SQL Api calls to Scala functions

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.

*

†with subtitle note
‡unity id: 200255928
§unity id: 200262103

NCSU, FALL 2018, October 18, 2018, Raleigh, NC, USA
2018.

3 Related Work

To the best of our knowledge after performing an exhaustive search, there is no automatic transformer for this problem. Although there are some good practices and tutorials on how to convert RDD API code to Dataset and Dataframe API, but those have to be done by a programmer and is a manual, time consuming process.

4 Methodology

5 Testing Criteria

6 Workload Distribution

7 Timeline

References

A Appendix

Text of appendix ...