FP&A with Spreadsheets and Spark

Oscar Castañeda-Villagrán
Universidad del Valle de Guatemala

#SAISEco2
About

• Researcher at Universidad del Valle de Guatemala.

• Research Interests:
  • Program Transformation,
  • Model-driven Data Product Design & Development
About

- Researcher at Universidad del Valle de Guatemala.

- Research Interests:
  - Program Transformation,
  - Model-driven Data Product Design & Development
About

- Researcher at Universidad del Valle de Guatemala.

- Research Interests:
  - Program Transformation,
  - Model-driven Data Product Design & Development
    - Design and Development of
About

• Researcher at Universidad del Valle de Guatemala.

• Research Interests:
  • Program Transformation,
  • Model-driven Design & Development
    ▸ Design and Development of
    ▸ Large-scale Data Products
About

- Researcher at Universidad del Valle de Guatemala.

- Research Interests:
  - Program Transformation,
  - Model-driven Data Product Design & Development
    - Design and Development of
    - Large-scale Data Products
    - with Spreadsheets as Models
Prototyping ...
Prototyping Spark programs with ...
Spreadsheets!
Agenda

- Problem Statement and Motivation
  - Architecture
- Program Transformation
  - Pipeline
  - Code-to-Code Transformation
- Code Generation
  - Abstract Tree
  - Parse Tree
- Spreadsheets as a DSL
  - Generating Code
- Demo
- Q&A
Disclaimer(s)

- Ongoing research ...

- FP&A is one use case, but Spreadsheets are much broader!
Disclaimer(s)

- Ongoing research …

- FP&A is one use case, but Spreadsheets are much broader!
  - E.g. People have even modeled Turing machines with Spreadsheets! [1]
Prototype FP&A programs using Spreadsheet formulas and automatically translate to Scala / Spark.
Problem Statement

Prototype *Any* program using Spreadsheet formulas and automatically translate to Scala / Spark.
Motivation

- At Spark Summit Europe 2016 I presented the **Sparksheet** code generator for Spreadsheet formulas.

- Initially **Sparksheet** supported only 5 Spreadsheet formulas, now it supports 150+ Spreadsheet formulas!

- Motivation is finding use cases.
Motivation

Automatically translate Spreadsheet datasets* to Spark data pipelines on Scala/Spark

*Spreadsheet dataset = Structured data + Spreadsheet formulas
Architecture
Architecture
Program Transformation

“A program transformation is any operation that takes a computer program and generates another program.”

https://en.wikipedia.org/wiki/Program_transformation
Program Transformation Pipeline
Demo #1

1. Show Spreadsheet model
2. Show Complex Spreadsheet Formula

http://bit.ly/2e0TZAl
Source program in Excel formula language.
Demo #2
Demo #2
Demo #2

Sparksheet

Complex

Spreadsheet Formula

Spark
Code-to-Code Transformation
Code-to-Code Transformation

“The input to the code generator typically consists of a parse tree or an abstract syntax tree.”

http://matplotlib.org

Generating Code

“An elegant way to generate code from an AST is to write a class for each non-terminal node in the tree, and then each node in the tree simply generates the piece of code that it is responsible for.”

http://www.codeproject.com/Articles/26975/Writing-Your-First-Domain-Specific-Language-Part
Generating Code

A practical way to generate code is to take a Parse Tree and write a pretty printer for the target language.
Generating Code (Example)

\[ \text{SUM}(A,C) \]
Generating Code (Example)

```
import org.apache.spark.sql._
import org.apache.spark.sql.functions._

def SUM(A: Int, C: Int): Int = {
    return A + C
}
val applySUM = udf(SUM _)
val sumDF = baseDF.withColumn("",
    applySUM(      
        col("A"),
        col("C")
    ))
```
Generating Code (Example)

\[ \text{SUM(A,C)} \]
Demo #3
Demo #3
Demo #3
Demo #3
Demo #3
Demo #3
Spreadsheets as a DSL

- Spreadsheet is a powerful data modeling tool.
- Start simple and evolve into a complex ML pipeline.
- Spreadsheets are suitable to many domains (FP&A is one such domain).
What have we seen?

- Spreadsheet applications as Prototypes for Spark programs
- Program Transformation
  - How to model as Pipeline
  - Why considered Code-to-Code Transformation
- How to Generate Code
  - AST (elegant)
  - Parse Tree (practical)
- Spreadsheets as a DSL
  - Generating Code
- Next Steps
Next Steps

• Use cases!

• Modeling Machine Learning in a Spreadsheet

• Prototype D|’s and ML|’s in a Spreadsheet
References

• A Grammar for Spreadsheet Formulas Evaluated on Two Large Datasets – Efthimia Aivaloglou, David Hoepelman & Felienne Hermans, Proceedings of SCAM ’15

• http://www.felienne.com/archives/2974

• Pictures in presentation from Boards of Canada video “roygbiv”

https://youtu.be/yT0gRc2c2wQ
Q&A
THANK YOU.
Email: ofcastaneda@uvg.edu.gt
Twitter: @oscar_castaneda