Verifying Equivalence of Spark Programs

Shelly Grossman[1], Sara Cohen[2], Shachar Itzhaky[3], Noam Rinetzky[1], and Mooly Sagiv[1]

1 **School of Computer Science, Tel Aviv University, Tel Aviv, Israel**
{shellygr,maon,msagiv}@tau.ac.il
2 **School of Engineering and Computer Science, The Hebrew University of Jerusalem, Jerusalem, Israel**
sara@cs.huji.ac.il
3 **Computer Science, Massachusetts Institute of Technology, USA**
shachari@mit.edu

Abstract

In this paper, we present a novel approach for verifying the equivalence of Spark programs. Spark is a recent framework for large scale data processing. Such frameworks, intended for data-intensive operations, share many similarities with traditional database systems, but do not possess the same optimization tools existing for databases. Our goal is to enable such optimizations by first providing the necessary theoretical setting for the analysis of distributed data processing frameworks and their special properties. We model Spark as a programming language which imitates Relational Algebra queries in the bag semantics, with operations as *map*, analogous to *project*;*filter*, analogous to *select*; and cartesian product, as well as an *aggregate* operation (fold). In addition, map, filter, and aggregate operations are described using *User Defined Functions* (*UDF*) acting on the elements, enriching the RA-like queries. We show decidable equivalence testing procedure for programs without aggregate operations, based on a symbolic representation of the elements in the database. We also prove a criterion for the decidability of equivalence of programs with aggregate operations. For Spark programs with aggregate operations that satisfy the decidability criterion, we present a sound and complete algorithm for verifying equivalence.

1998 ACM Subject Classification D.2.4 Software Engineering Software/Program Verification. D.1.3 Programming Techniques Concurrent Programming. F.3.2 Logics and Meanings of Programs Semantics of Programming Languages (Program analysis). H.2.3 Query languages. H.2.4 Distributed databases.

Keywords and phrases Spark, Map reduce, Program equivalence

Digital Object Identifier 10.4230/LIPIcs...

1 Introduction

The rise of Cloud computing and Big Data in the last decade allowed the advant of new programming models, with the intention of simplifying the development process for large-scale needs, letting the programmer focus on business-logic and separating it from the technical details of data management over computer clusters, distribution, communication and parallelization. The first model was MapReduce [8], It allowed programmers to define their logic by composing several iterations of map and reduce operations, where the programmer provided the required map and reduce functions in each step, and the framework was responsible for facilitating the dataflow to the provided functions. MapReduce gave a powerful, yet a clean and abstract programming model. Later on, other frameworks were developed, allowing programmers to write procedural code while still retaining the ability to

run it on a large computer cluster, without having the programmer to handle distribution and error recovery. One such framework is Apache Spark [16], in which programmers keep writing code in their programming language of choice, (e.g., Scala [4], Java [1], Python [3], or R [12]) but utilize a special object provided by Spark, called *resilient distributed dataset* (*RDD*), providing access to the distributed data itself and to perform transformations on it, using the cloud resources for actual computing.

The architecture of Spark comprises of a single master node, referred to as the *driver*, and *worker* nodes in a clustered computer environment. All the nodes have access to the program code, but the driver orchestrates its execution using the underlying Spark framework, abstracting away communications, error recovery, distribution, data partitioning, and parallelization. In particular, datasets are distributed, meaning that fragments of it, referred to as *partitions*, reside in the files system of some or all nodes [9]. The access to the data is via the *RDD* API. The *RDD* can be thought of as a simple database table, but which provides support for *User Defined Functions* (*UDF*), greatly increasing its expressive power.

Spark programs handle a family of common tasks involving large datasets, such as log parsing, database queries, training algorithms and different numeric computations in various fields. Due to this, many Spark programs share several properties: they are mostly short and relatively simple to read and understand. We believe that thanks to this nature of Spark programs, the problem of verifying a program's properties, or even program equivalence as we focus on in this paper, may become feasible, even decidable in a usable class of Spark programs.

Main Results. In this paper we define a simple programming language called $SPARK$, in which operations on a single RDD are abstracted as composite simply typed λ-calculus expressions. The operations correspond to operations in relational algebra, with additional aggregate operations. We describe the problem of *program equivalence* (*PE*), and provide a classification of $SPARK$ programs according to the decidability of the *PE* problem for programs in the same class. When *PE* is decidable, we show an algorithm for solving it.

Overview. Section 2 provides necessary preliminaries for the rest of the paper. In section 3 we give a complete formalization (syntax and semantics) of $SPARK$. In section 4.1 we describe the *PE* problem. In section 4 we discuss the most basic class of $SPARK$ programs, which are programs without aggregate expressions. We continue in Section 5, where we discuss different classes of $SPARK$ containing an aggregated expression.

2 Notations

This section provides necessary notations used throughout this paper.

Basic notations We denote the set of natural numbers (including zero) by $\mathbb{N}$, and by $\mathbb{N}^+$ the positive natural numbers. We also denote $\mathbb{Z}$ for integers (positive and non-positive), and $\mathbb{B}$ for booleans: tt, ff. The *undefined* value is denoted by $\perp$. We denote the *size* (number of elements) of a set X by $|X|$. We denote a general if-then-else operator (*ite*) as $ite(cond, then\_expr, else\_expr) = \begin{cases} then\_expr & cond \\ else\_expr & otherwise \end{cases}$. We denote a variable x belonging to type τ with $x{:}\tau$.

Tuples Let $X = X_1 \times \cdots \times X_n$ be a tuple domain of arity n. $p_i(X) = X_i$ is defined for $i \in \{1, \ldots, n\}$. Respectively, for an element $x = (x_1, \ldots, x_n) \in X$, $p_i(x) = x_i$ for $i \in \{1, \ldots, n\}$. The common arithmetic operations $(+, -, *)$ over integers are lifted to (vectorial) operations tuples in a point-wise manner.

Binary Relations Let $R \subseteq X \times Y$ be a binary relation. We use the following notations: **Projection ($i \in \{1, 2\}$):** $p_i(R) = \{p_i(t) \mid t \in R\}$, **Relation restriction:** $\sigma_\varphi(R) = \{t \in R \mid t \in \varphi\}$ where φ is a set, **Relation restriction by first element:** $\sigma_Z(R) = \sigma_{\{t \mid p_1(t) \in Z\}}(R)$ where Z is a set, $\sigma_x(R) = \sigma_{\{t \mid p_1(t) \in \{x\}\}}(R)$,

Functions A function f from X to Y is a relation between X and Y which is single valued $(\forall x \in X. |\sigma_x(f)| \leq 1)$. The domain of f is $\mathrm{dom}(f) = X$. We write $y = f(x)$ to denote that $(x, y) \in f$. We use $\cdot \hookrightarrow \cdot$ and $\cdot \to \cdot$ to denote *partial* and *total* functions, respectively. We denote the *support* $\mathrm{sup}(f) = p_1(f)$. We say that a function f is *undefined* at x if $x \notin \mathrm{sup}(f)$, and denote it by $f(x) = \perp$. If the support of f is empty then we write $f = \perp$.

Bags (multisets) A *bag* (*multiset*) m over a domain X is a partial function from X to $\mathbb{N}^+$. For $x \in \mathrm{dom}(m)$, we say that $x \in X$ is *in* m ($x \in m$) if $m(x) \neq \perp$, and refer to $m(x)$ as its *multiplicity* in m. Conversely, $x \in X$ is *not in* m ($x \notin m$) if $m(x) = \perp$. We write $m = \{\!\{Z\}\!\}$, where $Z \subseteq X$, to denote that Z is the support of the bag m. To define a bag m we write, similarly to set-comprehension notation, $\{\!\{x\,;\, m(x) \mid x \in \phi\}\!\}$, where ϕ is a set. Elements x for which $x \notin \phi$, do not belong in the bag ($m(x) = \perp$).

3 The *SPARK* language

In this section, we define the syntax of *SPARK*, a simple imperative programming language which allows to use Spark's *resilient distributed datasets* (*RDDs*) [16].

3.1 Data Model

Basic types. *SPARK* supports two primitive types: integers (`Int`) and booleans (`Boolean`). On top of this, the user can define types which are cartesian products of primitive types. In the following we use $\boldsymbol{c}$ to range over integer numerals (constants), $\boldsymbol{b} \in \{\texttt{true}, \texttt{false}\}$ to range over boolean constants, and τ to range over basic types and record types.

RDDs. In addition, *SPARK* allows the user to define *RDD*s. RDDs are bags of *records*, all of the same type. Hence, $RDD_{\boldsymbol{\tau}}$ denotes bags containing records of type $\boldsymbol{\tau}$.

Semantic Domains. We interpret the integers and boolean primitive types as *integers* ($\mathbb{Z}$) and *booleans* ($\mathbb{B}$), respectively. The interpretation of both primitive types is denoted $\mathrm{T} = \mathbb{Z} \cup \mathbb{B}$. The interpretation of all possible types (including mixed cartesian products of the primitive types) is denoted by $\mathcal{T} = \bigcup_n \mathrm{T}^n$.

The *RDD* type is interpreted as a *bag*. Therefore, $\boldsymbol{r}$ ranging over $RDD_{\boldsymbol{\tau}}$ is interpreted as a bag of type $[\![\boldsymbol{\tau}]\!]$, $[\![r]\!] = r \in ([\![\boldsymbol{\tau}]\!] \hookrightarrow \mathbb{N})$. We let $RDD = \bigcup_{\tau \in \mathcal{T}} RDD_\tau$, the semantic domain of RDDs over all possible record types $\tau \in \mathcal{T}$.

Interpretation operator We use $[\![\cdot]\!]$ (semantic brackets) to denote the mathematical interpretation of an expression. We shall see in the next subsections that it may be a function of an *environment* when the expressions contain variables. The exact meaning of environments and semantic interpretation of syntactic strings under environments is fully explained in Section 3.4.

3.2 Functional Model

Operations *RDD*s are analogous to database tables and as such the methods to query the *RDD*s are inspired by both *Relational Algebra* (*RA*) [] and Spark. *RA* has 5 basic operators, which are *Select*, *Project*, *Cartesian Product*, *Union* and *Subtract*. This paper focuses on the first three operators. The *Select* operator is analogous to *filter* in *SPARK*, and *Project* is analogous to *map*. The expressive power of *SPARK*'s *map* and *filter* is greater than their analogous *RA* operations thanks to the UDFs (see next), which allow *extended projection* as well as greater flexibility in executing complex operations on elements of different types.

UDFs. A special property of Spark is in allowing some of its standard operations to be *higher order functions* - to receive a function and to apply it on an RDD in a method defined by the operation. For example, we can *fold* an RDD containing integer numbers by providing a sum function of two integers to the *fold* transform. Such a function is called a *"User-Defined Function"*, or simply *UDF*, for short. The *signature* of a UDF contains information on the return type and the arguments types, and when applied in the context of an RDD operation, the signature should match both the RDD type and the operation on it (see typing rules in appendix 1). The syntax of the *body* of a UDF is the same as that of first-order simply typed lambda expressions []. For example: $sumMod10\colon \texttt{Int} \times \texttt{Int} \to \texttt{Int}$ is the signature of the following function: $sumMod10 = \lambda x, y.x\%10 + y\%10$. The types are omitted from the body of the function for brevity, but when the types are not clear from the context, we may also write it as: $sumMod10 = \lambda x\colon \texttt{Int}, y\colon \texttt{Int}.x\%10 + y\%10\colon \texttt{Int}$. Note that $sumMod10$ has two arguments, but we could also write it as a function with single argument having the following signature: $sumPairMod10\colon (\texttt{Int} \times \texttt{Int}) \to \texttt{Int}$, with body: $sumPairMod10 = \lambda(x, y).x\%10 + y\%10$.

We allow the definition of these functions to be parametric, meaning that there are free variables in the lambda expressions, which we wrap with an additional λ. For example, we could define a function that adds 1 to an integer: $\boldsymbol{f} = \lambda x\colon \texttt{Int}.x + 1$, but we could also make it more generic and flexible by writing $\boldsymbol{g} = \lambda a\colon \texttt{Int}.\lambda x\colon \texttt{Int}.x + a$. This is an example of a *parametric function*. Parametric functions can be transformed to regular functions by beta-reducing the first lambda abstraction: $\boldsymbol{g}(1)$ which is identical to $\boldsymbol{f}$.

3.3 Syntax

The syntax of *SPARK* language is defined in Figure 2.

Let: $isOdd = \lambda x\colon \texttt{Int}.\neg x\%2 = 0$
$doubleAndAdd = \lambda c\colon \texttt{Int}.\lambda x\colon \texttt{Int}.2 * x + c$
$sumFlatPair = \lambda A\colon \texttt{Int}, (x, y)\colon \texttt{Int} \times \texttt{Int}.A + x + y$

	$\mathrm{P1}(R_0\colon RDD_{\texttt{Int}}, R_1\colon RDD_{\texttt{Int}})$:
1	A = `filter`$(isOdd)(R_0)$
2	B = `map`$(doubleAndAdd(1))(A)$
3	C = `cartesian`(B, R_1)
4	v = `fold`$(0, sumFlatPair)(C)$
5	`return` v

Figure 1 Example *SPARK* program

Syntactic Categories We assume variables to be an infinite syntactic category, ranged over by $\boldsymbol{v}, \boldsymbol{b}, \boldsymbol{r} \in \texttt{Vars}$. Expressions range over $\boldsymbol{e}$. An integer constant is denoted c. Operations are divided to 4 categories: Relational, Mapping, Grouping, and Aggregating. Some of the operations require arguments, which may be either a primitive expression, an *RDD*, or a function. Functions range over $\boldsymbol{f}, \boldsymbol{F} \in \texttt{LambdaExpressions}$. Parametric functions are denoted by capital meta-variables ($\boldsymbol{F}$ as opposed to $\boldsymbol{f}$ for regular functions) and must always be given the list of parameters when passed to an operation.

Program structure The header of a program contains function definitions. Loops are not allowed in the body of a program. Variable declarations are in *SSA* (*Static Single Assignment*) form [7]. Variables are immutable by this construction. Programs have no side effects, do not change the inputs, and always return a value. The *program signature* will consist of its name, its input types and return type: $\boldsymbol{P}(\overline{\mathcal{T}_i}, \overline{RDD_i})\colon \mathcal{T}_o$

Example program Consider the example *SPARK* program in Figure 1.

From the example program we can see the general structure of *SPARK* programs: First, the functions that are used as UDFs in the program are declared and defined: $isOdd, sumFlatPair$ defined as *Fdef*, and $doubleAndAdd$ defined as a *PFdef*. The name of the program ($\boldsymbol{P} = P1$) is announced with a list of input RDDs (R_0, R_1) (*Prog* rule). Instead of writing *Let* l_1 *in Let* l_2 *in* ..., we use syntactic sugar, where each line of code contains a single '*Let*' definition, and the target expression (which may be a '*Let*' expression itself) follows. Then, 3 variables of RDD type are defined (A, B, C) and one integer variable (v). We can see in the definition of A an application of the *filter* operation, accepting the RDD R_0 and the function $isOdd$. For B's definition we apply the *map* operation with a parametric function $doubleAndAdd$ with the parameter 1, or simply the function $\lambda x.2 * x + 1$. C is the cartesian product of B and input RDD R_1. We apply an aggregation using *fold* on the RDD C, with an initial value 0 and the function $sumFlatPair$, which 'flattens' elements of tuples in C by taking their sum, and summing all those vectorial sums to a single value stored in the variable v. As we omit '*Let*' expressions from the example, we use the `return` keyword to return a value. The returned value is the integer variable v. The program's signature is $P1(RDD_{\texttt{Int}}, RDD_{\texttt{Int}})\colon \texttt{Int}$.

Basic Types	τ	::=	$\texttt{int} \mid \texttt{bool} \mid \tau \times \ldots \times \tau$
RDDs	RDD	::=	$\texttt{RDD}_\tau$
Variables	x	::=	$\boldsymbol{v} \mid \boldsymbol{r}$
Arithmetic Exp.	ae	::=	$c \mid ae + ae \mid -ae \mid c * ae \mid ae \,/\, c \mid ae \,\%\, c$
Boolean Exp.	be	::=	$\texttt{true} \mid \texttt{false} \mid e = e \mid ae < ae \mid \neg be \mid be \wedge be \mid be \vee be$
General Basic Exp.	e	::=	$ae \mid be \mid \boldsymbol{v} \mid (e\,, e) \mid \boldsymbol{p}_i(e) \mid \texttt{if}\ (b)\ \texttt{then}\ e\ \texttt{else}\ e$
Functions	$Fdef$	::=	$\texttt{def}\ \boldsymbol{f}\ =\ \lambda\overline{\boldsymbol{y}:\tau}\ \ e:\tau$
Parametric Functions	$PFdef$	::=	$\texttt{def}\ \boldsymbol{F}\ =\ \lambda\overline{\boldsymbol{x}:\tau}.\lambda\overline{\boldsymbol{y}:\tau}\ \ e:\tau$
RDD Exp.	re	::=	$\texttt{cartesian}(\boldsymbol{r},\boldsymbol{r}) \mid \texttt{map}(\boldsymbol{f})(\boldsymbol{r}) \mid \texttt{filter}(\boldsymbol{f})(\boldsymbol{r})$ $\mid \texttt{foldByKey}(\boldsymbol{e},\boldsymbol{f})(\boldsymbol{r})$
RDD Aggregation Exp.	ge	::=	$\texttt{fold}(\boldsymbol{e},\boldsymbol{f})(\boldsymbol{r})$
General Exp.	η	::=	$e \mid re \mid ge$
Program Body	E	::=	$Let\ \boldsymbol{x} = \eta\ in\ E \mid \eta$
Program	$Prog$	::=	$\boldsymbol{P}(\overline{\boldsymbol{r}:RDD_\tau}, \overline{\boldsymbol{v}:\tau})\ =\ \overline{Fdef}\quad \overline{PFdef}\quad E$

Figure 2 Syntax for $SPARK$

Constants	$[\![c]\!](\rho)$	=	c
Variables	$[\![\boldsymbol{v}]\!](\rho)$	=	$\rho(v)$
Unary operations	$[\![\texttt{uOp}\, e]\!](\rho)$	=	$uOp\,[\![e]\!](\rho)$
Binary operations	$[\![e\,\texttt{binOp}\,e]\!](\rho)$	=	$[\![e]\!](\rho)\,binOp\,[\![e]\!](\rho)$
Ternary operations	$[\![\texttt{if}\ e\ \texttt{then}\ e\ \texttt{else}\ e]\!](\rho)$	=	$ite\,([\![e]\!](\rho), [\![e]\!](\rho), [\![e]\!](\rho))$

Figure 3 Basic Expression semantics (for e) - `uOp`, `binOp`, and `terOp` are taken from Figure 2 : $\texttt{uOp} \in \{-, \neg, \boldsymbol{\pi}_i\}$, $\texttt{binOp} \in \{+, *, /, \%, =, <, \wedge, \vee, (,)\}$

3.4 Semantics

Program Environment We define a unified semantic domain $\mathcal{D} = \mathcal{T} \cup RDD$ for all types in $SPARK$. The *program environment* type:

$$\mathcal{E} = \texttt{Vars} \to \mathcal{D}$$

is a mapping from each variable in `Vars` to its value, according to type. A variable's type does not change during the program's run, nor does its value.

Data flow The *environment function* $\rho \in \mathcal{E}$ denotes the environment of the program. The environment function is initialized as $\rho = \perp$, and filled for input variables according to the inputs: $\boldsymbol{v}_1^{in}, \ldots, \boldsymbol{v}_k^{in}$ at initial environment: $\rho(\boldsymbol{v}_j^{in}) = [\![\boldsymbol{v}_j^{in}]\!]$, where $[\![\cdot]\!]$ is used to express the interpretation of the input in the semantic domain. We denote $[\![\boldsymbol{v}]\!](\rho) = \rho(\boldsymbol{v})$ as the *interpreted value* of the variable $\boldsymbol{v}$ of type $\mathcal{D}$. The semantics of expressions are straightforward and we provide the semantics with the current environment ρ (using $[\![\cdot]\!](\cdot)$), see Figure 3 for details. In Figure 4 we specify the behavior of $[\![\cdot]\!](\cdot)$ on the body of the program.

Program	$[\![Prog]\!]$	=	$[\![E]\!](\rho)$
Assignment	$[\![\boldsymbol{x} = \eta]\!](\rho)$	=	$\rho[\boldsymbol{x} \mapsto [\![\eta]\!](\rho)]$
Sequence	$[\![Let\ \ \boldsymbol{x} = \eta\ \ in\ E]\!](\rho)$	=	$[\![E]\!]([\![\boldsymbol{x} = \eta]\!](\rho))$

Figure 4 Semantics for $SPARK$ programs in structural induction. Semantics of η are described in Figures 3 and 6

Simple functions	$[\![\boldsymbol{f}]\!]$	=	$\lambda\overline{y{:}\tau}.\ [\![e]\!]$
Parametric functions	$[\![\boldsymbol{F}]\!]$	=	$\lambda\overline{x{:}\tau}, \overline{y{:}\tau}.\ [\![e]\!]$
Parametric functions with parameters	$[\![\boldsymbol{F}(\overline{e_p})]\!]$	=	$\lambda\overline{y{:}\tau}.\ [\![e[\overline{\boldsymbol{x}} \mapsto \overline{e_p}]]\!]$

Figure 5 Semantics for functions

Function and UDF semantics For UDFs, which are based on a restricted fragment of the *simply typed lambda calculus* [], we assume the syntax and semantics are the same as in the λ-calculus. Note however that the syntax does not allow passing higher order functions as UDFs, and forces any higher order function to be reduced to a first-order function beforehand. In addition, all parameters passed to UDF which are based on higher-order functions are read-only. The semantics of functions are defined in Figure 5.

Semantics of operations In Figure 6 the semantics of all RDD expressions, including aggregation, are explicitly stated.

Notes In the *bag semantics*:

- In *map*, if several elements y map to x by f, then the multiplicity of x is the sum of multiplicities of all y elements. In other words, if an element x appears n times, we apply the map UDF on it n times.
- fold *is well defined:* It should be noted that the Spark specification requires UDFs passed to aggregate operations to be **associative** for the value to be uniquely defined. By the assumed commutativity of $\boldsymbol{f}$, the order in which elements from the bag are chosen in the *fold* operation does not affect the result.
- *foldByKey* returns an RDD where each *key* k (defined to be the first element of the RDD type tuple) appears once, and the value is replaced with the fold result of all values v_i such that that were mapped to the same key in the original RDD: $(k, v_i) \in r$. Thus, we have a fold operator acting on the restriction of r to some key k, projected to the values.

Example For the example program Figure 1, suppose we were given the following input: $R_0 = \{\!\{(1;7),(2;1)\}\!\}, R_1 = \{\!\{(3;4),(5;2)\}\!\}$. Then: $\rho(A) = \{\!\{(1;7)\}\!\}$, $\rho(B) = \{\!\{(3;7)\}\!\}$, $\rho(C) = \{\!\{((3,3);28),((3,5);14)\}\!\}$, $\rho(v) = 28 * (3+3) + 14 * (3+5)$ and the program returns $\rho(v) = 280$.

4 Basic decidability results for $SPARK$

4.1 The *PE* problem definition

Next we describe the main decision problem which we try to solve for different classes of $SPARK$ programs.

$$
\begin{array}{lllll}
\textbf{Application:} & [\![\texttt{op(args)(rdds)}]\!] & = & [\![\texttt{op}]\!]([\![\texttt{args}]\!])([\![\texttt{rdds}]\!]) & \\
\textbf{Map:} & [\![\texttt{map}]\!](f)(r) & = & \{\!\{p_i(x)\,; \sum_{y \in r \wedge p_i(y)=p_i(x)} r(y) \mid x \in r\}\!\} & \\
\textbf{Filter:} & [\![\texttt{filter}]\!](b)(r) & = & \sigma_{\{x \mid b(x)\}}(r) & \\
\textbf{Cartesian:} & [\![\texttt{cartesian}]\!](r_1, r_2) & = & \{\!\{(x_1, x_2)\,; r_1(x) \cdot r_2(x) \mid x_1 \in r_1 \wedge x_2 \in r_2\}\!\} & \\
\textbf{Fold:} & [\![\texttt{fold}]\!](a_0, f)(r) & = & \begin{cases} f([\![\texttt{fold}]\!](a_0, f)(r'), a) & r = r' \cup \{\!\{a; 1\}\!\},\ \textit{where } a \in r \\ v & r = \bot \end{cases} & \\
\textbf{FoldByKey:} & [\![\texttt{foldByKey}]\!](a_0, f)(r) & = & \{\!\{(k, v); 1 \mid (k, \_) \in r \wedge v = [\![\texttt{fold}]\!](a_0, f)(\pi_2(r \restriction_{(k,\_)}))\}\!\} &
\end{array}
$$

Figure 6 Semantics for RDD expressions (re, ge)

$$
\begin{array}{lllll}
\textbf{Program:} & \phi_P(E, k) & = & \begin{cases} (t_{E'}[x \mapsto t_\eta], m),\ \textit{where} & E = \textit{Let } x = \eta \textit{ in } E' \\ (t_{E'}, n) = \phi_P(E', k), (t_\eta, m) = \phi_P(\eta, n) & \\ \phi_P(\eta, k) & E = \eta \end{cases} & \\
\textbf{Basic exps.:} & \phi_P(e, k) & = & (e, k) & \\
\textbf{RDD exprs.:} & & & & \\
\textbf{Map:} & \phi_P(\texttt{map}(f)(r), k) & = & (f(t), m),\ \textit{where } (t, m) = \phi_P(r, k) & \\
\textbf{Filter:} & \phi_P(\texttt{filter}(f)(r)) & = & \textit{ite}(f(t) = tt, (t, m), \bot),\ \textit{where } (t, m) = \phi_P(r, k) & \\
\textbf{Cartesian:} & \phi_P(\texttt{cartesian}(r_1, r_2), k) & = & ((t_{r_1}, t_{r_2}), m),\ \textit{where } (t_{r_1}, n) = \phi_P(r_1, k), (t_{r_2}, m) = \phi_P(r_2, n) &
\end{array}
$$

$$
\textbf{Input RDDs:}\quad \phi_P(r, k) = \begin{cases} ((p_1(\mathbf{x}_r^{(k)}), \ldots, p_n(\mathbf{x}_r^{(k)}), k+1) & r \in \overline{r} \\ (r, k) & \textit{otherwise} \end{cases}
$$

$$
\begin{array}{c}
\textit{Let } P : \boldsymbol{P}(\overline{\boldsymbol{r}}, \overline{\boldsymbol{v}}) = \overline{\boldsymbol{F}}\,\overline{\boldsymbol{f}}\,E \\
\Phi(P) = t,\ \textit{where } \phi_P(E, 0) = (t, \_)
\end{array}
$$

Figure 7 Compiling $SPARK$ to logical terms (ϕ).

The Program Equivalence (*PE*) problem Let P_1 and P_2 be $SPARK$ programs, with signature $P_i(\overline{T}, \overline{RDD_T}) : \tau$ for $i \in \{1, 2\}$. We use $[\![P_i]\!]([\![\overline{v}]\!], [\![\overline{r}]\!])$ to denote the result of P_i. We say that P_1 and P_2 are *equivalent*, if for all input values $\overline{v}$ and RDDs $\overline{r}$, it holds that $[\![P_1]\!]([\![\overline{v}]\!], [\![\overline{r}]\!]) = [\![P_2]\!]([\![\overline{v}]\!], [\![\overline{r}]\!])$.

4.2 An equivalency checking framework

Introduction and intuition We will describe an alternative, equivalent semantics for $SPARK$ where the program is interpreted as a term in first order logic over the theory of integer arithmetic. This term is called the *program term* and denoted $\Phi(P)$ for program P, and its exact specification appears in Figure 7. The variables of the term are taken from the input RDDs. We take for example the program in Figure 1. The Φ function is defined using the function ϕ whose purpose is to maintain consistent variable renaming. ϕ is applied recursively on the expression returned by the program. We simplify by running

ϕ_{P1} on each line of the program, top-down:

$$\begin{aligned}
\phi_P(A,0) &= \phi_P(\texttt{filter}(isOdd)(R_0),0) = (ite(isOdd(t),t,\bot),n \text{ where } (t,n) = \phi_P(R_0,0)\\
&=^{\phi_P(R_0,0)=\mathbf{x}_{R_0}^{(0)},1} ite(isOdd(\mathbf{x}_{R_0}^{(0)}),\mathbf{x}_{R_0}^{(0)},\bot),1\\
\phi_P(B,1) &= \phi_P(doubleAndAdd(1)(A),1) = doubleAndAdd(1)(A),1\\
&= doubleAndAdd(1)(ite(isOdd(\mathbf{x}_{R_0}^{(0)}),\mathbf{x}_{R_0}^{(0)},\bot)),1\\
\phi_P(C,1) &= \phi_P(\texttt{cartesian}(B,R_1),1) =^{\phi_P(B,1)=B,1}= (B,p_1(\phi_P(R_1,1))),p_2(\phi_P(R_1,1))\\
&=^{\phi_P(R_1,1)=\mathbf{x}_{R_1}^{(1)},2} (B,\mathbf{x}_{R_1}^{(1)}),2\\
\phi_P(v,2) &= \phi_P(\texttt{fold}(0,sumFlatPair)(C),2) =^{\phi_P(C,2)=C,2} [C]_{0,sumFlatPair},2\\
\Phi(P) &= p_1(\phi_P(v,2)) = [C]_{0,sumFlatPair} = [(B,\mathbf{x}_{R_1}^{(1)})]_{0,sumFlatPair}\\
&= [(doubleAndAdd(1)(ite(isOdd(\mathbf{x}_{R_0}^{(0)}),\mathbf{x}_{R_0}^{(0)},\bot)),\mathbf{x}_{R_1}^{(1)})]_{0,sumFlatPair}
\end{aligned}$$

Representative elements of RDDs To create the program terms, we use variables that are based on the input RDDs. Such variables are called *representative elements*. In a program that receives an input RDD r^{in}, we denote the representative element of r^{in} as: $\mathbf{x}_{r^{in}}$. The set of possible valuations of that variable is equal to the bag defined by r^{in}, and an additional 'undefined' value ($\bot$), for the empty RDD. Therefore $\mathbf{x}_{r^{in}}$ ranges over $\text{dom}(r^{in}) \cup \{\bot\}$. By abuse of notation, the term for an RDD computed in a $SPARK$ program is also called a representative element.

Matching representative elements For two program terms to be comparable, they must depend on the same input RDDs. For example:

	P1(R_0: $RDD_{\texttt{Int}}$, R_1: $RDD_{\texttt{Int}}$):	P2(R_0: $RDD_{\texttt{Int}}$, R_1: $RDD_{\texttt{Int}}$):
1	`return map`$(\lambda x.1)(R_0)$	`return map`$(\lambda x.1)(R_1)$

We see that $P1$ and $P2$ have the same program term (1), but the multiplicity of that element in the output bag is different and depends on the source input RDD. In $P1$, its multiplicity is the same as the size of R_0, and in $P2$ it's the same as the size of R_1. $P1$ and $P2$ are therefore not equivalent, because we can provide inputs R_0, R_1 of different sizes.
For each program term $\Phi(P)$ we therefore consider $FV(\Phi(P))$, the *set of free variables*. Each free variable has some source input RDD, and an input RDD may have more than one free variable representing it in the program term. In the example, the set of free variables of $P1$ is $FV(\Phi(P1)) = \{\mathbf{x}_{R_0}\}$, and of $P2$ it is $FV(\Phi(P1)) = \{\mathbf{x}_{R_1}\}$. For programs to be equivalent, there must be an isomorphism between the sets, mapping each free variable to a single variable with the same source input RDD.

Compiling $SPARK$ to logical terms using representative elements Let P be a $SPARK$ program. We use standard notations $\vec{r^{in}}$ for the inputs of P, and r^{out} for the output of P. The term $\Phi(P)$ is called the *program term of* P and it is defined by structural induction according to the notations in Figure 7. We write the set of free variables of the program term $FV(\Phi(P))$ as a vector: $(\mathbf{x}_{r^{in}_{j_k}})_{k=1}^{n_P}$, where n_P is the number of free variables. A vector of valuations to the free variables is denoted $\vec{x} = (x_1,\ldots,x_n)$ and satisfies $x_k \in r^{in}_{j_k}$ for $k \in \{1,\ldots,n\}$. The *SR semantics* (*Symbolic Representation Semantics*) of a program that returns an RDD-type output is the bag that is obtained from all possible valuations to the free variables:

$$SR(P)(\vec{r^{in}}) = \{\!\{\Phi(P)[FV(\Phi(P)) \mapsto \vec{x}] \mid \Phi(P)[FV(\Phi(P)) \mapsto \vec{x}] \neq \bot \wedge \forall k \in \{1,\ldots,n_P\}.x_k \in r^{in}_{j_k}\}\!\}$$

Assigning a concrete valuation to the free variables of $\Phi(P)$ returns an element in the output RDD r^{out}. By taking all possible valuations to the term with elements from $\vec{r^{in}}$, we get the bag equal r^{out}. We elaborate later on programs that return a basic type by using aggregate operations on RDDs.

▸ **Proposition 4.1.** *Let $P : \boldsymbol{P}(\overline{\boldsymbol{r}}) = \overline{\boldsymbol{F}}\,\overline{\boldsymbol{f}}\,E$ be a SPARK program, $[\![P]\!]$ be the interpretation of its output according to the defined semantics, and $SR(P)$ by the symbolic representation semantics of P as described earlier. Without loss of generality, we do not consider programs with non-RDD inputs. Then, for any input $\vec{r^{in}}$, we have:*

$$SR(P)(\vec{r^{in}}) = [\![P]\!]([\![\vec{r^{in}}]\!])$$

Proof. See appendix 2. ◂

Program equivalence problem formalization using representative elements Given two programs P, Q receiving as input a series of RDDs $\vec{r^{in}} = (r_1^{in}, \ldots, r_n^{in})$. We assume w.l.o.g. the programs are non-trivial, meaning they do not return the empty RDD for any choice of inputs. We define isomorphism of sets of free variables for P, Q as an injective and onto mapping: $\mathcal{S} : FV(\Phi(P)) \xrightarrow{\sim} FV(\Phi(Q))$ such that $\forall k \in \{1, \ldots, n\}. S(\mathbf{x}^{(k)}_{r^{in}_{j_k}}) = \mathbf{x}^{(i)}_{r^{in}_{j_i}} \wedge j_k = j_i$

The *PE* problem becomes the problem of proving the following:

$$(*) \quad FV(\Phi(P)) \simeq^{\mathcal{S}} FV(\Phi(Q))$$
$$(**) \quad \neg(\exists \vec{x}\ \Phi(P)[FV(\Phi(P)) \mapsto \vec{x}] \neq \Phi(Q)[\mathcal{S}(FV(\Phi(P))) \mapsto \vec{x}])$$

where the choice of $\vec{r^{in}}$ is arbitrary, and $\mathcal{S}$ is non-deterministically chosen from all legal isomorphisms.

4.3 Proof technique

4.3.1 Equivalency without RDDs

The following proposition follows directly from the decidability of Presburger arithmetic [].

▸ **Proposition 4.2.** *Given two SPARK programs P and Q which use only integer or boolean basic types and no RDDs,* PE *is decidable.*

Proof. *sketch.* Let the signatures of P and Q be $P(\overline{T}){:}\tau$, $Q(\overline{T}){:}\tau$. Expression and functions in *SPARK* belong to an extension of the Presburger arithmetic to integer numbers, which is decidable [5]. If the return types τ are tuples, then on per-element basis, equivalency is decidable: For each element of the returned tuple we get an equation of functions applied to the variables in the programs, definable in the Presburger arithmetic, so the problem of program equivalence reduces to solving the Presburger formula. It is decidable by using a decision procedure such as *Cooper's algorithm* [6]. [1] ◂

[1] A remark on complexity: Cooper's algorithm has an upper bound of $2^{2^{2^{pn}}}$ for some $p > 0$ and where n is the number of symbols in the formula [15]. In practice, our experiments show that Cooper's algorithm on non-trivial formulas returns almost instantly, even on commodity hardware.

1. Compare the program signatures to see input types match. If not, return **not equivalent**.
2. Using typing rules (see appendix 1), check if the output RDDs' types match. If the types do not match, return **not equivalent**.
3. Apply for both P and Q the SR() semantics by building the *program terms* (denoted $\Phi(P), \Phi(Q)$). Construction of the program terms is done by structural induction according to the rules appearing in Figure 7.
4. Verify that:
$$FV(\Phi(P)) = FV(\Phi(Q))$$
If not, output **not equivalent**.
5. **a.** Choose an isomorphism $\mathcal{S}$ of the representative elements of the input RDDs in both P, Q.
   b. We check the following formula is satisfiable:
$$\exists \vec{v}. \Phi(P)[FV(\Phi(P)) \mapsto \vec{v}] \neq \Phi(Q)[S(FV(\Phi(Q))) \mapsto \vec{v}]$$
   c. If it is satisfiable, go back to (a) and repeat until finding an unsatisfiable formula, or all possible isomorphisms were exhausted.
   d. If the formula is unsatisfiable, return **equivalent**.
   e. If all isomorphisms were exhausted without finding an unsatisfiable formula, then return **not equivalent**.

Figure 8 An algorithm for solving PE

4.3.2 Basic operations: Map and Filter

▸ **Lemma 1** (Decidability for basic programs: *map* , *filter* and *cartesian* operations)**.** *Given two SPARK programs P and Q which use only basic types coming from T (where all integer expressions are expressible using the* Presburger arithmetic*), only* map *,* filter *and* cartesian *are allowed operations for RDDs,* PE *is decidable.*

Proof. For non-RDD return types we already proved in Proposition 4.2 - without aggregate operations, basic types can not be influenced by operations performed on RDDs. Thus, we can remove all RDD operations from the program and get an equivalent program in the setting of Proposition 4.2. For RDDs we provide an algorithm in Figure 8, which is a decision procedure. The correctness of the algorithm follows from the equivalency of the SR() semantics and the semantics defined in 3.4. It checks that two SR() expressions are equal, and the process terminates:

- Termination follows from having no loops in $SPARK$, and a finite number of RDDs giving a finite number of isomorphisms to check.
- Each step involving verifying a formula is decidable because the language of $SPARK$ is limited to expressions in the decidable Presburger arithmetic theory.

◂

Note: All examples use syntactic sugar for '*Let*' expressions. For brevity, instead of applying ϕ on the underlying '*Let*' expressions, we apply it line-by-line from the top-down. Finding the isomorphism $\mathcal{S}$ between variable names in both programs is done automatically in all programs, but formally it is part of the decision procedure. In addition, we assume that in programs returning an RDD-type, the RDD is named r^{out}, and the programs always end with `return` r^{out}. Thus, $\Phi(P) = \phi_P(r^{out})$.

▸ Example 4.1 (Basic optimization - operator pushback). This example shows a common optimization of pushing the filter/selection operator backward, to decrease the size of the dataset.

	P1(R: RDD_{Int}):	P2(R: RDD_{Int}):
1	$R' = \texttt{map}(\lambda x.2 * x)(R)$	$R' = \texttt{filter}(\lambda x.x < 7)(R)$
2	`return` $\texttt{filter}(\lambda x.x < 14)(R')$	`return` $\texttt{map}(\lambda x.x + x)(R')$

RDD return type: Both programs return an RDD of integers.

Free variables: $FV(\Phi(P1)) = \{\mathbf{x}_R\} = FV(\Phi(P2))$. Thus, the sets of free variables are equal.

Analysis of representative elements:

$\phi_{P1}(R') = 2 * \mathbf{x}_R$, and $\phi_{P1}(r^{out}) = \begin{cases} \varphi & \varphi < 14 \wedge \varphi = \phi_{P1}(R') \\ \bot & otherwise \end{cases} = \begin{cases} 2 * \mathbf{x}_R & 2 * \mathbf{x}_R < 14 \\ \bot & otherwise \end{cases}$.

$\phi_{P1}(R') = \begin{cases} \mathbf{x}_R & \mathbf{x}_R < 7 \\ \bot & otherwise \end{cases}$, and $\phi_{P2}(r^{out}) = (\lambda x.x + x)(\phi_{P1}(R')) = \begin{cases} \mathbf{x}_R & \mathbf{x}_R < 7 \\ \bot & otherwise \end{cases} + \begin{cases} \mathbf{x}_R & \mathbf{x}_R < 7 \\ \bot & otherwise \end{cases}$.

We need to verify that:

$$\forall \mathbf{x}_R. \begin{cases} 2 * \mathbf{x}_R & 2 * \mathbf{x}_R < 14 \\ \bot & otherwise \end{cases} = \begin{cases} \mathbf{x}_R & \mathbf{x}_R < 7 \\ \bot & otherwise \end{cases} + \begin{cases} \mathbf{x}_R & \mathbf{x}_R < 7 \\ \bot & otherwise \end{cases}$$

To prove this, we need to encode the cased expressions in Presburger arithmetic. Undefined ($\bot$) values indicate 'don't care' and are not part of the Presburger arithmetic. However, they can be handled by assuming the 'if' condition is satisfied, and verifying that the condition is indeed satisfied equally for all inputs. The first condition, therefore, is that both 'if' conditions agree on all possible values. The second condition is that the resulting expressions are equivalent.

▸ Proposition 4.3 (Schemes for converting conditionals to a normal form). *We write a series of universally true schemes for translating the* filter *cased expression to Presburger arithemtic when appearing in an equivalence formula:*

1. *The following useful identity for applying functions on a conditional is true:*

$$f\left(\begin{cases} e & cond \\ \bot & otherwise \end{cases}\right) = \begin{cases} f(e) & cond \\ \bot & otherwise \end{cases}$$

2. *Equivalence of functions of conditionals:*

$$f\left(\begin{cases} e & cond \\ \bot & otherwise \end{cases}\right) = g\left(\begin{cases} e' & cond' \\ \bot & otherwise \end{cases}\right) \iff (cond \iff cond') \wedge (cond \implies f(e) = g(e'))$$

3. *Equivalence of a function of a conditional and an arbitrary expression:*

$$f\left(\begin{cases} e & cond \\ \bot & otherwise \end{cases}\right) = e' \iff cond \wedge f(e) = e'$$

The Fold operator:	$[\varphi]_{init,f}$
The Fold By Key operator:	$\langle\varphi\rangle_{init,f}$

Figure 10 Aggregate operators for *SPARK* programs with aggregations

4. *Applying a function with multiple arguments on multiple conditionals (a function receiving* $\perp$ *input as one of its arguments returns a* $\perp$*):*

$$f\left(\begin{cases} e & cond \\ \perp & otherwise \end{cases}, \begin{cases} e' & cond' \\ \perp & otherwise \end{cases}\right) = \begin{cases} f(e,e') & cond \wedge cond' \\ \perp & otherwise \end{cases},$$

5. *Applying a function with multiple arguments on a conditional and a general expression:*

$$f\left(\begin{cases} e & cond \\ \perp & otherwise \end{cases}, e'\right) = \begin{cases} f(e,e') & cond \\ \perp & otherwise \end{cases}$$

6. *The two last rules define the base case for functions with more than 2 arguments where at least one of the arguments is a conditional.*
7. *Unnseting of nested conditionals*

$$\begin{cases} \begin{cases} e & c_{int} \\ \perp & otherwise \end{cases} & c_{ext} \\ \perp & otherwise \end{cases} = \begin{cases} e & c_{int} \wedge c_{ext} \\ \perp & otherwise \end{cases}$$

Using Cooper's algorithm and the above schemes, we can prove the equivalence formula is true. See Figure 9 for an implementation. ■

```
integer_ qelim
<<forall x.
(2*x<14 <=> x<7) /\ (2*x<14 ==> 2*x = x+x)>>;;
- :  fol formula = <<true>>
```

Figure 9 Output of Cooper ML implementation proving this example

For additional examples, refer to appendix 3.

▸ **Theorem 2.** *The algorithm described in Figure 8 is a decision procedure for the* PE *problem in SPARK programs without aggregations (*fold ,foldByKey *).*

5 Aggregate expressions

In the following section we discuss how the existing framework can be extended to prove equivalence of *SPARK* programs containing aggregate expressions.

Extending SR(P) with aggregate expressions The terms for aggregate operations are given using special operators (Figure 10). We extend the definition of ϕ accordingly in Figure 11. The $[\varphi]_{.,.}$ operator *binds* all variables in φ. The $\langle\varphi\rangle_{.,.}$ operator is syntactic sugar for $(p_1(\varphi), [p_2(\varphi)]_{.,.})$ which *binds* all variables in $p_2(\varphi)$ which are not in $p_1(\varphi)$. That is, the variables of $p_1(\varphi)$ are still free in the fold expression. The motivation is explained in 5.5.

$$\begin{array}{lrcl}
\textbf{Fold:} & \phi_P([\![\texttt{fold}]\!](f,e)(r),k) & = & ([t]_{e,f},m),\ \textit{where}\ (t,m)=\phi_P(r,k) \\
\textbf{FoldByKey:} & \phi_P([\![\texttt{foldByKey}]\!](f,e)(r),k) & = & (\langle t\rangle_{e,f},m),\ \textit{where}\ (t,m)=\phi_P(r,k)
\end{array}$$

Figure 11 Compiling $SPARK$ to logical terms for aggregate operations.

5.1 Single aggregate as the final expression

The simplest case of programs in which an aggregation operator appears, is one where a single aggregate operation is performed and it is the last RDD operation.

▸ Example 5.1 (Maximum and minimum). Below is a simple example of 2 equivalent programs representing the simple case of single aggregate operation in the end of the program:

Let: $max = \lambda M, x.\texttt{if}(x > M)\ \texttt{then}\ \{x\}\ \texttt{else}\ \{M\}$
$min = \lambda M, x.\texttt{if}(x < M)\ \texttt{then}\ \{x\}\ \texttt{else}\ \{M\}$

	P1($R : RDD_{\texttt{Int}}$):	P2($R : RDD_{\texttt{Int}}$):
1	`return` $\texttt{fold}(\bot, max)(R)$	$R' = \texttt{map}(\lambda x.0 - x)(R)$
2		`return` $0\text{-}\texttt{fold}(\bot, min)(R')$

In the above example we compute the maximum element of a numeric RDD in two different methods, in the first program by getting the maximum directly, and in the second by getting the additive inverse of the minimum of the additive inverses of the elements. The equivalence formula is:

$$[\mathbf{x}_R]_{\bot,max} = 0 - [0 - \mathbf{x}_R]_{\bot,min} = -[-\mathbf{x}_R]_{\bot,min}$$

To prove that the two reduced results are equal we use an inductive claim:

$$\forall x, A, A'.A = -A' \Rightarrow max(A,x) = -min(A',-x)$$

$$max(A,x) \overset{?}{=} -min(A',-x)$$

$$\begin{cases} A & A > x \\ x & otherwise \end{cases} \overset{?}{=} -\begin{cases} A' & A' < -x \\ -x & otherwise \end{cases} = \begin{cases} -A' & A' < -x \\ x & otherwise \end{cases} \overset{A=-A'}{=} \begin{cases} A & -A < -x \\ x & otherwise \end{cases} = \begin{cases} A & A > x \\ x & otherwise \end{cases}$$

Indeed, by replacing $A' = -A$ we get equal expressions. ■

5.1.1 Basic proof method

The inductive claim is generalized for the class of programs discussed here, of programs performing a single *aggregate* operation after a series of *map*, *filter* and *cartesian* operations (without self-products, that is without variable renaming). Those definitions lead to the following lemma:

▸ **Lemma 3.** *Let $R_0 \in RDD_{\sigma_0}, R_1 \in RDD_{\sigma_1}$, and denote their representative elements φ_0, φ_1 respectively. We assume φ_0, φ_1 were composed from* map, filter *and cartesian product (without self products). Let there be two* fold *functions $f_0 : \xi_0 \times \sigma_0 \to \xi_0, f_1 : \xi_1 \times \sigma_1 \to \xi_1$, two* initial *values $init_0 : \xi_0, init_1 : \xi_1$, and two functions $g : \xi_0 \to \xi, g' : \xi_1 \to \xi$. We have: if*

$$FV(\varphi_0) \simeq FV(\varphi_1),\ \textit{denoted}\ FV \tag{1}$$

$$g(init_0) = g'(init_1) \tag{2}$$

$$\forall \vec{v}, A_{\varphi_0} : \xi_0, A_{\varphi_1} : \xi_1 . g(A_{\varphi_0}) = g'(A_{\varphi_1}) \implies g(f_0(A_{\varphi_0}, \varphi_0[FV \mapsto \vec{v}])) = g'(f_1(A_{\varphi_1}, \varphi_1[FV \mapsto \vec{v}])) \tag{3}$$

then $g([\varphi_0]_{init_0,f_0}) = g'([\varphi_1]_{init_1,f_1})$

Induction length and relation to size of bags The proof of lemma 3 assumes the sets of free variables to be isomorphic, otherwise the induction termination is not well defined. One possibility is that the size of participating RDDs may not be equal. As a workaround, suppose we take a valuation from the union of FVs, using $\perp$ values if needed. Let there be two equal fold expressions under lemma's 3 premises, and a a valuation $\vec{v}$ in the valuation sequence returning a non-$\perp$ value in R_0, and $\perp$ valuation to R_1. We get:

$$\begin{array}{rrrcl} g(A_{\varphi_0,i}) = g'(A_{\varphi_1,i}) & \wedge & g(f_0(A_{\varphi_0,i}, \varphi_0[FV(\varphi_0) \mapsto \vec{v}])) & = & g'(f_1(A_{\varphi_1,i}, \perp)) = g'(A_{\varphi_1,i}) \\ & \Longrightarrow & g(f_0(A_{\varphi_0,i}, \varphi_0[FV(\varphi_0) \mapsto \vec{v}])) & = & g(A_{\varphi_0,i}) \\ \Longrightarrow^{\varphi_0[FV(\varphi_0)\mapsto\vec{v}]\neq\perp} & & \forall A, c. f_0(A, c) & = & A \end{array}$$

In that case, we see in the last transition that the lemma's conditions are fulfilled only if the fold functions are constant, namely the intermediate value returned is never changed by it. However, fold UDFs which are constant have no application in practice, thus we ignore programs containing them.

Lemma 3 shows that an inductive proof of the equality of folded values is *sound*. The meaning is that given any two folded expressions which are not equivalent, the lemma always reports them as non-equivalent. We show a constraint on 3, for which the method is also complete.

Examples Refer to appendix 5.

5.1.2 Completeness

On its surface, the inductive claim does not permit a *sound* ***and complete*** method of verifying the equivalence for the restricted class of programs we defined. However, if at least one of the transformations applied on the aggregated expression is an injection, the equivalence of the programs implies the inductive claim, making it a *complete* proof method. The underlying principle allowing it, is that if we presumed the inductive claim to be false while the equivalence of the programs is true, then we could trim the RDDs to a prefix of the same size that violates the inductive claim, which would mean a violation of the assumption on equivalence. We formulate this intuition in the next paragraphs.

We begin with showing that when we apply on folded expressions a non-injective function in both programs, the lemma may fail:

▸ Example 5.2 (Non-injective modification of folded expressions). Non-injective transformations can weaken the inductive claim, resulting in failure to prove it. As a result, lemma 3 fails to prove the equivalence of the following two programs.

	P1(R: $RDD_{\texttt{Int}}$):	P2(R: $RDD_{\texttt{Int}}$):
1	$R' = \texttt{map}(\lambda x.x\%3)(R)$	$R' = \texttt{fold}(0, \lambda A, x.A + x)(R)$
2	`return` $\texttt{fold}(0, \lambda A, x.(A + x)\%3)(R') = 0$	`return` $R'\%3 = 0$

To prove the equivalence, we should check by induction the equality of both boolean results. Taking $g(x) = \lambda x.x = 0$, $g'(x) = \lambda x.(x \ mod \ 3) = 0$ and attempt to prove by induction the following claim:

$$[x \ mod \ 3]_{0,+ \ mod \ 3} = 0 \Leftrightarrow [x]_{0,+ \ mod \ 3} \ mod \ 3 = 0$$
$$\forall x, A, A'.A = 0 \iff A' \ mod \ 3 = 0 \implies (A + x \ mod \ 3) \ mod \ 3 = 0 \iff (A' + x) \ mod \ 3 = 0$$

fails. To illustrate, suppose that in the induction hypothesis we have $A = 1, A' = 2$. Then the hypothesis that says $A = 0 \iff A' \ mod \ 3 = 0$ is satisfied, but it cannot be said that

$(A+x \ mod\ 3)\ mod\ 3 = 0 \iff (A'+x)\ mod\ 3 = 0$ (take $x = 1$: $((1+(1\%3))\%3 = 2, (2+1)\%3 = 0)$). And indeed, Cooper's algorithm outputs `'false'` for it:

```
integer_qelim
<<forall x,a,b.
(a = 0 <=> (exists w.  (b-3*w) = 0))
==>
( (exists y.  ((a+x-3*y) = 0)) <=> (exists z.  ((b+x-3*z) = 0) ) ) >>;;
- :  fol formula = <<false>>
```

Figure 12 Output of Cooper ML implementation showing the inductive claim is false here.

From the above example we derive the following lemma:

▸ **Lemma 4.** *Under the premises of lemma 3, if we have:*

$$(*) \quad g([\varphi_0]_{init_0,f_0}) = g'([\varphi_1]_{init_1,f_1}))$$

and in addition, g is injective or g' is injective, then we can prove $(*)$ by induction.

Proof. Suppose w.l.o.g g is injective. Then we attempt to prove the following by induction, which is equivalent to $(*)$:

$$(**) \quad [\varphi_0]_{init_0,f_0} = g^{-1}(g'([\varphi_1]_{init_1,f_1}))$$

Case 1: induction base not satisfied: Assume $init_0 \neq g^{-1}(g(init_1))$. We take R_0, R_1 to be empty bags. Then $[\varphi_j]_{init_j,f_j} = init_j$ for $j \in \{0,1\}$, and from $(**)$, we get: $init_0 = g^{-1}(g(init_1))$, contradiction.

Case 2: induction step not satisfied: Assume $init_0 = g^{-1}(g(init_1))$, and:

$$\exists \vec{v}, A_{\varphi_0}, A_{\varphi_1}.A_{\varphi_0} = g^{-1}(g(A_{\varphi_1})) \wedge f_0(A_{\varphi_0}, \varphi_0[FV(\varphi_0) \mapsto \vec{v}]) \neq g^{-1}(g'(f_1(A_{\varphi_1}, \varphi_1[FV(\varphi_1) \mapsto \vec{v}]))$$

Let the sequence of valuations $\langle \vec{a_1}, \ldots, \vec{a_n} \rangle$ be the generators of the intermediate values $A_{\varphi_0}, A_{\varphi_1}$. We take RDDs R_0, R_1 defined as follows for $j \in \{0,1\}$:

$$R_j = \bigcup_{i=1,\ldots,n} \{\!\{\varphi_j[FV(\varphi_j) \mapsto \vec{a_i}]\}\!\} \cup \{\!\{\varphi_j[FV(\varphi_j) \mapsto \vec{v}]\}\!\}$$

For this choice of RDDs we have:

$$[\varphi_j]_{init_j,f_j} = f_j(A_{\varphi_j}, \varphi_j[FV(\varphi_j) \mapsto \vec{v}])$$

But, from the assumption:

$$f_0(A_{\varphi_0}, \varphi_0[FV(\varphi_0) \mapsto \vec{v}]) \neq g^{-1}(g'(f_1(A_{\varphi_1}, \varphi_1[FV(\varphi_1) \mapsto \vec{v}]))$$

we get :

$$[\varphi_0]_{init_0,f_0} \neq g^{-1}(g'([\varphi_1]_{init_1,f_1}))$$

contradiction. ◂

Lemma 4 shows that the inductive process can be applied on injective mappings of folded expressions in order to prove their equivalence.

5.2 Multiple aggregates

Another relatively simple case of *SPARK* programs is when the program contains multiple aggregate operations, but they are independent of each other. Namely, the result of one aggregate operations is not used in the other one, and a final result can be proven by a set of formulas, each of which is dependent only on one aggregated result from each of the tested programs.

▸ Example 5.3 (Independent fold). Below are 2 programs which return a tuple containing the sum of positive elements in its first element, and the sum of negative elements in the second element. We show that by applying lemma 14 on each element of the resulting tuple, we are able to show the equivalency.

$$\text{Let:}\quad h:(\lambda(P,N),x.\begin{cases}(P+x,N) & x\geq 0\\ (P,N-x) & otherwise\end{cases}$$

	P1(R: RDD_{Int}):	P2(R: RDD_{Int}):
1	`return fold`$((0,0),h)(R)$	R_P = `filter`$(\lambda x.x\geq 0)(R)$
2		R_N = `map`$(\lambda x.-x)$(`filter`$(\lambda x.x<0)(R)$)
3		p = `fold`$(0,\lambda A,x.A+x)(R_P)$
4		$n=-$`fold`$(0,\lambda A,x.A+x)(R_N)$
5		`return` (p,n)

w.l.o.g. we show the application of lemma 14 to second element of the tuple. First,

$$\phi_{P2}(R_N)=\begin{cases}-\mathbf{x}_R & \mathbf{x}_R<0\\ \bot & otherwise\end{cases}$$

We let $g=\lambda x.x$ and $g'=\lambda x.-x$. Induction base case is obvious. Induction step:

$$\forall x,(P,N),N'.N'=N \implies p_2(h((P,N),x))=N'+\begin{cases}-x & x<0\\ \bot & otherwise\end{cases}$$

Substituting for $p_2\circ h$, we get that we need to prove:

$$\begin{cases}N & x\geq 0\\ N-x & otherwise(x<0)\end{cases} \overset{?}{=} N'+\begin{cases}-x & x<0\\ \bot & otherwise\end{cases}$$

$$\overset{\bot \text{ is neutral}}{=} \begin{cases}N'-x & x<0\\ N' & otherwise\end{cases}$$

$$\overset{N=N'}{=} \begin{cases}N-x & x<0\\ N & otherwise\end{cases}$$

$$= \begin{cases}N-x & x<0\\ N & x\geq 0\end{cases}$$

And the equivalence follows. ■

▸ **Lemma 5.** *Let $R_1\in RDD_{\sigma_1},\ldots,R_k\in RDD_{\sigma_k}$, and denote the representative elements φ_i for $i\in\{1,\ldots,k\}$. We assume the φ_i are based only on* map *and* filter *operations and* cartesian *without self-products. Let there be k* fold *UDFs $f_i:\xi_i\times\sigma_i\to\xi_i$, and k initial values $init_i$:ξ_i. Let a similar set of RDDs, representative elements,* fold *UDFs and initial values, with all denotions having a $'$. Let there be 2 functions g:$\xi_1\times\cdots\times\xi_k\to\xi$, g':$\xi'_i\times\cdots\times\xi'_{k'}\to\xi$.*

We denote a vector $\vec{v} = FV(\varphi_1, \ldots, \varphi_k, \varphi'_1, \ldots, \varphi'_{k'})$ *of the free variables in all representative elements. We have: if*

$$g(init_1, \ldots, init_k) = g'(init'_1, \ldots, init'_{k'}) \tag{1}$$

$$\begin{aligned} &\forall \vec{v}, A_{\varphi_1}{:}\xi_1, \ldots, A_{\varphi_k}{:}\xi_k, A_{\varphi'_1}{:}\xi'_1, \ldots, A_{\varphi'_{k'}}{:}\xi'_{k'}.g(A_{\varphi_1}, \ldots, A_{\varphi_k}) = g'(A_{\varphi'_1}, \ldots, A_{\varphi'_{k'}}) \implies \\ &g(f_1(A_{\varphi_1}, \varphi_1[FV(\varphi_1) \mapsto \vec{v}]), \ldots, f_k(A_{\varphi_k}, \varphi_k[FV(\varphi_k) \mapsto \vec{v}])) = \\ &g'(f'_1(A_{\varphi'_1}, \varphi'_1[FV(\varphi'_1) \mapsto \vec{v}]), \ldots, f'_k(A_{\varphi'_{k'}}, \varphi'_{k'}[FV(\varphi'_{k'}) \mapsto \vec{v}])) \end{aligned} \tag{2}$$

then $g([\varphi_1]_{init_1,f_1}, \ldots, [\varphi_k]_{init_k,f_k}) = g'([\varphi'_1]_{init'_1,f'_1}, \ldots, [\varphi'_{k'}]_{init'_{k'},f'_{k'}})$

5.3 A class for which *PE* is undecidabile

We show a reduction of Hilbert's 10'th problem to PE. We assume towards a contradiction that PE is decidable under the premises of lemma 5 with represntative elements based also on the *cartesian* operation. Let there be a polynom p over k variables $x_1, \ldots, x_k$, and coefficients $a_1, \ldots, a_k$. For each variable x_i we assume the existence of some RDD R_i with x_i elements. We use $SPARK$ operations and the input RDDs R_i to represent the value of the polynom P for some valuation of the x_i. For each summand in the polynom p, we define a translation φ:

- $x_i \longrightarrow^{\varphi} [\texttt{map}(\lambda x.1)(R_i)]_{0,+}$
- $x_i x_j \longrightarrow^{\varphi} [\texttt{cartesian}(\texttt{map}(\lambda x.1)(R_i), \texttt{map}(\lambda x.1)(R_j))]_{0,+}$
- By induction, $x_i^2 \longrightarrow^{\varphi} [\texttt{cartesian}(\texttt{map}(\lambda x.1)(R_i), \texttt{map}(\lambda x.1)(R_i))]_{0,+}$.
  This rule as well as the rest of the powers follow according to the previous rule. For a degree k monom, we apply the *cartesian* operation k times.
- ${x_i}^0 \longrightarrow^{\varphi} 1$, trivially.
- $am(x) \longrightarrow^{\varphi} a\varphi(m(x))$ where $m(x)$ is a monomial with coeffeicient 1 of the variable x, thus we have already defined φ for it.
- $aq(x_{i_1}, \ldots, x_{i_j}) \longrightarrow^{\varphi} a\varphi(q(x_{i_1}, \ldots, x_{i_j}))$ where $q(x)$ is a monomial with coefficient 1 and multiple variables for which φ was defined in the previous rules.
- $\varphi(p(a_1, \ldots, a_k; x_1, \ldots, x_k)) = \sum_{i=1}^{k} \varphi(a_i q(x_{i_1}, \ldots, x_{i_k})))$ follows by structural induction on the previous rules.

We generate the following instance of the PE problem:

	P1($R_1, \ldots, R_k$: $RDD_{\texttt{Int}}$):	P2($R_1, \ldots, R_k$: $RDD_{\texttt{Int}}$):
1	`return` $\varphi(p) \neq 0$	`return` tt

By choosing input RDDs of the cardianlity of R_i equal to the matching variable x_i we can simulate any valuation to the polynom p. If $P1$ returns true, then the valuation is not a root of the polynom p. Thus, if it is equivalent to the 'true program' $P2$, then the polynom p has no roots. Therefore, if the algorithm solving PE outputs 'equivalent' then the polynom p has no root, and if it outputs 'not equivalent' then the polynom p has some root, where $x_i = |[\![R_i]\!]|$. Thus we have polynomial reduction to Hilbert's 10'th problem.

▸ **Theorem 6** (Basic decidability for *PE*). *For two $SPARK$ programs Q_1, Q_2 returning a single or a tuple of RDDs, and for two $SPARK$ programs T_1, T_2 which act on the output of Q_1, Q_2 respectively by applying fold functions on them, we define $P_i = T_i \circ Q_i$. The* PE *problem is decidable if either T_1 or T_2 apply an injective transform on the folded expressions.*

Proof. A conclusion from lemmas 3, 4, 14, 5 ◂

5.4 Nested aggregations

We saw in 6 a proof of decidability for a fragment of $SPARK$ programs. In the following subsection we present more complex $SPARK$ programs on which the theorem applies, the method for proving the equivalence, and the cases on which it fails. Those programs have a value of an aggregate operation used in later aggregations (i.e. 'nested' aggregations). We see that the inductive method is sound in handling those cases, and that under certain conditions, it is complete too.

▸ Example 5.4 (Conditional summation). The following example takes the *sum* of all elements which are greater than the *count* of elements in an RDD.

Let: $f : (\lambda A, (a, b).A + b)$
$+ : \lambda A, x.A + x$

	P1(R: $RDD_{\texttt{Int}}$):	P2(R: $RDD_{\texttt{Int}}$):
1	$R' = \texttt{map}(\lambda x.(x, 2))(R)$	$R' = \texttt{map}(\lambda x.(x, 1))(R)$
2	$sz = \texttt{fold}(0, f)(R')$	$sz = \texttt{fold}(0, f)(R')$
3	$B = \texttt{filter}(\lambda x.x > sz)(R)$	$B = \texttt{filter}(\lambda x.x > 2 * sz)(R)$
3	$\texttt{return fold}(0, +)(B)$	$\texttt{return fold}(0, +)(B)$

The equivalence condition is:

$$\left[\begin{cases}\mathbf{x}_R & \mathbf{x}_R > \phi_{P1}(sz)\\ \bot & \text{otherwise}\end{cases}\right]_{0,+} = \left[\begin{cases}\mathbf{x}_R & \mathbf{x}_R > 2 * \phi_{P2}(sz)\\ \bot & \text{otherwise}\end{cases}\right]_{0,+}$$

Replacing sz, sz' we get:

$$\left[\begin{cases}\mathbf{x}_R & \mathbf{x}_R > [(\mathbf{x}_R^{(1)}, 2)]_{0,f}\\ \bot & \text{otherwise}\end{cases}\right]_{0,+} = \left[\begin{cases}\mathbf{x}_R & \mathbf{x}_R > 2 * [(\mathbf{x}_R^{(1)}, 1)]_{0,f}\\ \bot & \text{otherwise}\end{cases}\right]_{0,+}$$

After formally applying lemma 3 we get that the above is equivalent if and only if $\phi_{P1}(sz) = 2 * \phi_{P2}(sz)$, that is:

$$[(\mathbf{x}_R^{(1)}, 2)]_{0,f} = 2 * [(\mathbf{x}_R^{(1)}, 1)]_{0,f}$$

In this case, we set $g = \lambda x.x, g' = \lambda x.2 * x$, and get:

$$0 = g(0) = 2 * 0 \tag{1}$$

$$\begin{aligned}\forall x, A, A'.A = 2 * A' \implies f(A, (x, 2)) &= A + 2\\ &= 2 * A' + 2\\ &= 2 * (A' + 1)\\ &= f(A', (x, 1))\end{aligned} \tag{2}$$

Proving the equivalence. ■

This property is reflected in the following proposition:

▸ **Proposition 5.1.** *Let there be two $SPARK$ programs P_1, P_2 and returning aggregated expressions $[a_i(\vec{r})]_{init_i, f_i}$. Let there be two other $SPARK$ programs, T_1, T_2, also returning aggregated expressions, $[b_i(\vec{r'}, a)]_{init_T, g}$ (the aggregation function is equal in both T_1, T_2). PE is decidable for $T_1 \circ P_1, T_2 \circ P_2$ if and only if PE is decidable for P, P'.*

Proof. $\phi_{P_1} = [a_1(\vec{r})]_{init_1,f_1}, \phi_{P_2} = [a_2(\vec{r})]_{init_2,f_2}$ are the program terms of P_1, P_2, respectively. a, a' are terms without aggregations. Let $[b_i(\vec{r'}, \phi_{P_i})]_{init_T,g}$ for $i \in \{1,2\}$ be the program terms for $T_1 \circ P_1, T_2 \circ P_2$, respectively. The programs are equivalent if and only if the program terms are equivalent. As we have the same aggregation function on both representations, we can check the equivalence of $b_1(\vec{r'}, [a_1(\vec{r})]_{init_1,f_1}))$ and $b_2(\vec{r'}, [a_2(\vec{r})]_{init_2,f_2})$ instead. The decidability or undecidability is determined by corollary **??**. ◂

From proposition 5.1, it can be concluded that representative elements which are an injection as a function of the folded values have a decidable equivalence checking procedure. In the above example, the expression:

$$f(sz) = \lambda x. \begin{cases} x & x > sz \\ \bot & \text{otherwise} \end{cases}$$

is an injective function of sz - each such expression, when choosing a certain sz, is a different function of x.

This allows us to define an algorithm for verifying complex equivalences with aggregated queries. The idea is to apply nested inductive proofs (on the nested expressions) during the proof of the induction step of the outer aggregations.

▸ **Theorem 7.** *Let there be two SPARK programs P_1, P_2, returning an expression of the form: $f_i([\phi_i(\vec{x}, a_i(\vec{y}))]_{init_i,g_i})$, The* PE *problem is decidable if exists $i \in \{1,2\}$ such that the following expressions are all injective:*

1. f_i
2. $\mathcal{F} = \lambda A, \vec{x}. f_i(g_i(A_i, \phi_i(\vec{x}, a_i(\vec{y}))))$

Proof. We apply lemma 3 on the expressions returned by P_i:

$$f_1(init_1) = f_2(init_2) \tag{1}$$

$$\forall \vec{x}, A_1, A_2. f_1(A_1) = f_2(A_2) \implies f_1(g_1(A_1, \phi_1(\vec{x}, a_1(\vec{y})))) = f_2(g_2(A_2, \phi_2(\vec{x}, a_2(\vec{y})))) \tag{2}$$

a_1, a_2 are aggregated expressions: $a_i = [\psi_i(\vec{y})]_{j_i,h_i}$. So we apply lemma 3 again. For brevity, we denote: $\mathcal{F}' = f_i(g_i(A_i, \phi_i(\vec{x}, a_i(\vec{y})))) = \mathcal{F}(A_i, \vec{x})$

$$\mathcal{F}'(j_1) = \mathcal{F}'(j_2) \tag{1}$$

$$\forall \vec{y}, B_1, B_2. \mathcal{F}'(B_1) = \mathcal{F}'(B_2) \implies \mathcal{F}'(h_1(B_1, \psi_1(\vec{y}))) = \mathcal{F}'(h_2(B_2, \psi_2(\vec{y}))) \tag{2}$$

As we know that for at least one of the sides we have injective functions for both applications of lemma 3, then by theorem 6 we have a decision procedure for the equivalence. ◂

Therefore, by using the conditions determined by theorem 7, there is no need to assume the outer aggregation function is equal, as stated in this lemma:

▸ **Lemma 8.** *Let there be two SPARK programs P_1, P_2 and returning aggregated expressions $[a_i(\vec{r})]_{init_i,f_i}$. Let there be two other SPARK programs, T_1, T_2, also returning aggregated expressions, $[b_i(\vec{r'}, a)]_{init_{T_i},g_i}$.* PE *is decidable for $T \circ P, T' \circ P'$ if and only if* PE *is decidable for P, P'.*

Proof. $\phi_{P_1} = [a_1(\vec{r})]_{init_1,f_1}, \phi_{P_2} = [a_2(\vec{r})]_{init_2,f_2}$ are the program terms of P_1, P_2, respectively. a, a' are terms without aggregations. Let $[b_i(\vec{r'}, \phi_{P_i})]_{init_{T_i},g_i}$ for $i \in \{1,2\}$ be the program terms for $T_1 \circ P_1, T_2 \circ P_2$, respectively. The programs are equivalent if and only if the program terms are equivalent. By applying lemma 3 on the outer aggregation, we need to check:

$$g_1(init_{T_1}) = g_2(init_{T_2}) \tag{1}$$

$$\begin{aligned} &\forall \vec{x'}, A_1, A_2. g_1(A) = g_2(A_2) \implies \\ &g_1(f_1(A_1, b_1(\vec{r'}, \phi_P)[\vec{r'} \mapsto \vec{x'}])) = g_2(f_2(A_2, b_2(\vec{r'}, \phi_{P'})[\vec{r'} \mapsto \vec{x'}])) \end{aligned} \tag{2}$$

Verifying the second equation is decidable under the assumptions of theorem 7. ◂

Below are several examples:

▸ Example 5.5 (Increase all elements by count of elements and count). In this example we add the *count* of the RDD's elements to all elements, and count the elements of the result. Of course, this is the same as simply counting the elements.

Let: $count = \lambda A, x. A + 1$

	P1(R: RDD_{Int}):	P2(R: RDD_{Int}):
1	$c = \texttt{fold}(0, count)(R)$	$c = \texttt{fold}(0, count)(R)$
2	$R' = \texttt{map}(\lambda c. \lambda x. x + c)(R)$	`return` c
3	`return` $\texttt{fold}(0, count)(R')$	

Program term of $P1$: $\phi_{P1} = [\mathbf{x}_R + [\mathbf{x}_R]_{0,count}]_{0,count}$. Program term of $P2$: $\phi_{P2} = [\mathbf{x}_R]_{0,count}$. We apply lemma 8. Induction: base case is trivial. Step: Let $A = A'$. Need to prove $A + 1 = A' + 1$ which is immediate. *count* maps each element which is not $\perp$ to the intermediate value, so the value of $[\mathbf{x}_R]_{0,count}$ does not affect the result. ■

▸ Example 5.6 (Increase all elements by count of elements and sum). If the number of elements in the input RDD is n and their sum is s, then the following two programs calculate $s + n * n$.

$count = \lambda A, x. A + 1$
$sum = \lambda A, x. A + x$
Let: $count = \lambda A, x. A + 1$
$addConst = \lambda c. \lambda x. x + c$
$replaceWith = \lambda c. \lambda x. c$

	P1(R):	P2(R):
1	$c = \texttt{fold}(0, count)(R)$	$c = \texttt{fold}(0, count)(R)$
2	$R' = \texttt{map}(addConst(c))(R)$	$s = \texttt{fold}(0, sum)(R)$
3	`return` $\texttt{fold}(0, sum)(R')$	$R' = \texttt{map}(replaceWith(c))(R)$
4		`return` $\texttt{fold}(0, sum)(R') + s$

Program term of $P1$:

$$\phi_{P1} = [\mathbf{x}_R + [\mathbf{x}_R]_{0,count}]_{0,sum}$$

Program term of $P2$:

$$\phi_{P2} = [[\mathbf{x}_R]_{0,count}]_{0,sum} + [\mathbf{x}_R]_{0,sum}$$

We have both nested aggregations and multiple aggregate terms in $P2$. We apply 2 lemmas: lemma 5 and lemma 8. Base case: 0=0+0. Step: Let $A = B + C$. Need to prove: $A + (\mathbf{x}_R + [\mathbf{x}_R^{(1)}]_{0,count}) = B + [\mathbf{x}_R^{(1)}]_{0,count} + C + \mathbf{x}_R$. We get immediately that $[\mathbf{x}_R^{(1)}]_{0,count}$ cancel out of both sides of the equation, and it is enough to prove $A + \mathbf{x}_R = B + C + \mathbf{x}_R$, which is true by the induction hypothesis. ■

▸ Example 5.7 (Count number of elements bigger than half the maximal element). This example once again shows how sum and count can be equivalent, but the aggregations are applied on an RDD modified according to the result of a previous aggregation.

$$\text{Let:} \quad \begin{array}{c} max = \lambda A, x.ite(A < x, x, A) \\ sum_2 = \lambda A, (x, y).A + y \\ count = \lambda A, x.A + 1 \\ biggerThanHalfOf = \lambda m.\lambda x.2 * x > m \end{array}$$

	P1(R: $RDD_{\texttt{Int}}$):	P2(R: $RDD_{\texttt{Int}}$):
1	$m = \texttt{fold}(\bot, max)(R)$	$m = \texttt{fold}(\bot, max)(R)$
2	$R' = \texttt{filter}(biggerThanHalfOf(m))(R)$	$R' = \texttt{filter}(biggerThanHalfOf(m))(R)$
3	$R'' = \texttt{map}(\lambda x.(x, 1))(R')$	$\texttt{return fold}(0, count)(R')$
4	$\texttt{return fold}(0, sum_2)(R')$	

Program term of $P1$:

$$\phi_{P1} = \left[\begin{cases} (\mathbf{x}_R, 1) & 2 * \mathbf{x}_R > [\mathbf{x}_R^{(1)}]_{\bot, max} \\ \bot & otherwise \end{cases} \right]_{0, sum_2}$$

Program term of $P2$:

$$\phi_{P2} = \left[\begin{cases} \mathbf{x}_R & 2 * \mathbf{x}_R > [\mathbf{x}_R^{(1)}]_{\bot, max} \\ \bot & otherwise \end{cases} \right]_{0, count}$$

According to theorem 7, we can check the induction base case for the outer aggregation which is trivial: $0 = 0$, for the step we assume $A = A'$, and we need to prove:

$$\begin{cases} A + 1 & 2 * \mathbf{x}_R > [\mathbf{x}_R^{(1)}]_{\bot, max} \\ A & otherwise \end{cases} = \begin{cases} A' + 1 & 2 * \mathbf{x}_R^{(1)} > [\mathbf{x}_R^{(1)}]_{\bot, max} \\ A' & otherwise \end{cases}$$

Which is straightforward. ■

▸ Example 5.8 (Sum and count polynomial from two RDDs). In this example we use two RDDs.

$$\text{Let:} \quad \begin{array}{c} sum = \lambda A, x.A + x \\ count = \lambda A, x.A + 1 \\ addConst = \lambda c.\lambda x.x + c \\ replaceWith = \lambda c.\lambda x.c \end{array}$$

	P1(R_0: $RDD_{\texttt{Int}}$, R_1: $RDD_{\texttt{Int}}$):	P2(R_0: $RDD_{\texttt{Int}}$, R_1: $RDD_{\texttt{Int}}$):
1	$c = \texttt{fold}(0, count)(R_0)$	$c = \texttt{fold}(0, count)(R_0)$
2	$R' = \texttt{map}(addConst(c))(R_1)$	$s = \texttt{fold}(0, sum)(R_1)$
3	$\texttt{return fold}(0, sum)(R')$	$R' = \texttt{map}(replaceWith(c))(R)$
3		$\texttt{return fold}(0, sum)(R') + s$

Program term of $P1$:

$$\phi_{P1} = [\mathbf{x}_{R_1} + [\mathbf{x}_{R_0}]_{0, count}]_{0, sum}$$

Program term of $P2$:

$$\phi_{P2} = [[\mathbf{x}_{R_0}]_{0, count}]_{0, sum} + [\mathbf{x}_{R_1}]_{0, sum}$$

We apply theorem 7. Base case: 0=0+0. Step: Let x_0 a valuation for $\mathbf{x}_{R_0}$, and x_1 a valuation for $\mathbf{x}_{R_1}$. Assume $A = B + C$. Need to prove: $A + (x_1 + [\mathbf{x}_{R_0}]_{0, count}) = B + [\mathbf{x}_{R_0}]_{0, count} + C + x_1$. We get that it is enough to prove $A + x_1 = B + C + x_1$, x_0, which is true by the induction hypothesis. x_0 is not used. ■

5.5 By-key operations

SG: use new macros in the entire section

Previously we defined the fold by key operator $\langle x_r \rangle_{i,f}$ and the semantics of the *foldByKey* operation. According to the semantics, the folded value for each key is affected only by the set of values that match to the key in the bag. When the set of variables appearing in the term for the key and the set of variables appearing in the term for the value are disjoint, we have: $\langle x_r \rangle_{i,f} = (p_1(x_r), [p_2(x_r)])_{i,f}$. Otherwise, the term is more complex, and we shall illustrate it using the following example:

▸ Example 5.9 (Mixed key-value variables in value). In this example there is an RDD with dependencies between the key and the value:

Let: $f = \lambda A, x.A + x$
$f' = \lambda A, (x, y).A + y$

	P1(R: $RDD_{K\times V}$):	P2(R: $RDD_{K\times V}$):
1	`return foldByKey`$(0, f)(R)$	$R' =$ `map`$(\lambda(k, v).(k, (k, v)))(R)$
2		`return foldByKey`$(0, f')(R')$

We denote $\mathbf{x}_R = (\mathbf{x}_R^k, \mathbf{x}_R^v)$. For $P1$: $\phi_{P1}(r^{out}) = \langle(\mathbf{x}_R^k, \mathbf{x}_R^v)\rangle_{0,f} = (\mathbf{x}_R^k, [\mathbf{x}_R^v]_{0,f})$. For $P2$: $\phi_{P2}(r^{out}) = \langle(\mathbf{x}_R^k, (\mathbf{x}_R^k, \mathbf{x}_R^v)\rangle_{0,f'}$. Suppose we take some constant $\mathbf{x}_R^k$, and choose some $\mathbf{x}_R^v$ such that $(\mathbf{x}_R^k, \mathbf{x}_R^v) \in [\![R]\!]$. If so, we need to prove $[(\mathbf{x}_R^k, \mathbf{x}_R^v)]_{0,f'} = [\mathbf{x}_R^v]_{0,f}$ By induction: base case is trivial, for A, A' such that $A = A'$, and some arbitrary choice of $\mathbf{x}_R^v$, we have:

$$f'(A', (\mathbf{x}_R^k, \mathbf{x}_R^v)) = A' + \mathbf{x}_R^v = f(A', \mathbf{x}_R^v) = f(A, \mathbf{x}_R^v)$$

as required.

Specification of the *foldByKey* term: Let us denote $FV_K(\phi)$ the set of free variables of the key section of a representative element ϕ: $FV(p_1(\phi))$, and $FV_V(\phi)$ the set of free variables of the value section of a representative element ϕ: $FV(p_2(\phi))$. We denote a new function based on the original *foldByKey* UDF: $f_k = \lambda\vec{k}{:}T^{|FV_K(\phi)|}.\lambda\vec{v}{:}T^{|FV_V(\phi)\setminus FV_K(\phi)|}.f[FV_K(\phi) \mapsto \vec{k}]$

Then: $\langle\phi\rangle_{i,f} = (p_1(\phi), [p_2(\phi)]_{i,f_k(\vec{a})})$ where $\vec{a}{:}T^{|FV_K(\phi)|}$

The *fold* operator binds all free variables in the representative element on which the *fold* operator is applied. The *foldByKey* operator binds only the variables that appear in the value section of the representative element but not in the key-section. Thus, it is a representative element whose free variables are those that belong to the key-section of the representative element on which the *foldByKey* operator was applied.

For two *foldByKey* expressions, we consider first the methodology for testing two RDDs with independent key and value terms for equivalence.

▸ **Lemma 9.** *Let there be two programs P, P' with program terms $\phi_P(r^{out}) = (p_1(t), [p_2(t)]_{i,f})$ and $\phi_{P'}(r^{out}) = (p_1(t'), [p_2(t')]_{i',f'})$. We have $\phi_P(r^{out}) = \phi_{P'}(r^{out'})$ if and only if:*

- $p_1(t) = p_1(t')$ *and*
- $[p_2(t)]_{i,f} = [p_2(t')]_{i',f'}$

For non-independent key-value terms, the extension to lemma 9 is natural:

▸ **Lemma 10.** *Let there be two programs P, P' with program terms $\phi_P(r^{out}) = (p_1(t), [p_2(t)]_{i,f_k(\vec{a})})$ and $\phi_{P'}(r^{out}) = (p_1(t'), [p_2(t')]_{i',f'_k(\vec{a'})})$. We have $\phi_P(r^{out}) = \phi_{P'}(r^{out'})$ if and only if:*

- $p_1(t) = p_1(t')$ *and*

- $[p_2(t)]_{i,f_k(\vec{a})} = [p_2(t')]_{i',f'_k(\vec{a'})}$

Proof. Correctness is by definition. ◂

We can therefore extend previous results to aggregated results which are by-key.

▸ **Corollary 11.** *Theorem 6 can be applied to the aggregated value of a by-key expression once the equivalence of the keys was proven.*

▸ Example 5.10. Basic example:

Let: $sum = \lambda A, x.A + 2 * x$
$double_2 = \lambda(x, y).(x, 2 * y)$

	P1(R: $RDD_{\texttt{Int×Int}}$):	P2(R: $RDD_{\texttt{Int×Int}}$):
1	$R' = \texttt{foldByKey}(0, sum)(R)$	$R' = \texttt{map}(double_2)(R)$
2	$\texttt{return map}(double_2)(R')$	$\texttt{return foldByKey}(0, sum)(R')$

The representative elements and program term of $P1$ are:

$$\phi_{P1}(R') = (p_1(\mathbf{x}_R), [p_2(\mathbf{x}_R)]_{0,sum})$$

$$\phi_{P1}(r^{out})(p_1(\mathbf{x}_R), 2 * [p_2(\mathbf{x}_R)]_{0,sum})$$

The representative elements and program term of $P2$ are:

$$\phi_{P2}(R') = (p_1(\mathbf{x}_R, 2 * p_2(\mathbf{x}_R))$$

$$\phi_{P2}(r^{out})(p_1(\mathbf{x}_R), [2 * p_2(\mathbf{x}_R)]_{0,sum})$$

Need to prove, by lemma 9: $p_1(\mathbf{x}_R = p_1(\mathbf{x}_R$ (immediate), and

$$2 * [p_2(\mathbf{x}_R)]_{0,sum} = [2 * p_2(\mathbf{x}_R)]_{0,sum}$$

We proceed with an application of lemma 3:

$$\forall x = (x_1, x_2), A, A'.2 * A = A' \implies 2 * (A + x_2) = A' + (2 * x_2)$$

We have $2 * (A + x_2) = 2 * A + 2 * x_2 = A' + 2 * x_2$ as required. ■

▸ Example 5.11. This example shows how we can prove equivalence of by-key transformations even when we attempt to 'trick' the theory by making the definition of the keys 'fluid' - meaning keys with tuples of variant length.

Let: $f = \lambda A, (x_1, x_2).(max(p_1(A), x_1), min(p_2(A), x_2))$
$g = \lambda A, x.max(A, x)$

	P1(R: $RDD_{\texttt{Int×(Int×Int)}}$):	P2(R: $RDD_{\texttt{Int×(Int×Int)}}$):
1	$R_1 = \texttt{filter}(\lambda(x, (y, z)).x = y \wedge y = z)(R)$	$R_1 = \texttt{filter}(\lambda(x, (y, z)).x = y \wedge y = z)(R)$
2	$R_2 = \texttt{map}(\lambda(x, (y, z)).((x, y), z))(R_1)$	$\texttt{return foldByKey}((\perp, \perp), f)(R_1)$
3	$R_3 = \texttt{foldByKey}(\perp, g)(R_2)$	
4	$\texttt{return map}(\lambda((x, y), z).(x, (y, z)))(R_3)$	

SG: [originally, in the PODS98 paper, there was sum(y) instead of max(y). But this is not really equivalent when relations are allowed to contain bags, for example two instances of (1,1,1) in R]

The representative elements and program term of $P1$ are:

$$\phi_{P1}(R_1) = \begin{cases} \mathbf{x}_R & p_1(p_1(\mathbf{x}_R)) = p_1(p_2(\mathbf{x}_R)) \wedge p_1(p_2(\mathbf{x}_R)) = p_2(p_2(\mathbf{x}_R)) \\ \perp & otherwise \end{cases}$$

$$\phi_{P1}(R_2) = \begin{cases} ((p_1(p_1(\mathbf{x}_R)), p_1(p_2(\mathbf{x}_R))), p_2(p_2(\mathbf{x}_R))) & p_1(p_1(\mathbf{x}_R)) = p_1(p_2(\mathbf{x}_R)) \wedge p_1(p_2(\mathbf{x}_R)) = p_2(p_2(\mathbf{x}_R)) \\ \bot & otherwise \end{cases}$$

$$\phi_{P1}(R_3) = \begin{cases} ((p_1(p_1(\mathbf{x}_R)), p_1(p_2(\mathbf{x}_R))), [p_2(p_2(\mathbf{x}_R))]_{\bot,g}) & p_1(p_1(\mathbf{x}_R)) = p_1(p_2(\mathbf{x}_R)) \wedge p_1(p_2(\mathbf{x}_R)) = p_2(p_2(\mathbf{x}_R)) \\ \bot & otherwise \end{cases}$$

$$\phi_{P1}(r^{out}) = \begin{cases} (p_1(p_1(\mathbf{x}_R)), (p_1(p_2(\mathbf{x}_R)), [p_2(p_2(\mathbf{x}_R))]_{\bot,g})) & p_1(p_1(\mathbf{x}_R)) = p_1(p_2(\mathbf{x}_R)) \wedge p_1(p_2(\mathbf{x}_R)) = p_2(p_2(\mathbf{x}_R)) \\ \bot & otherwise \end{cases}$$

The representative elements and program term of $P2$ are:

$$\phi_{P2}(R_1) = \begin{cases} \mathbf{x}_R & p_1(p_1(\mathbf{x}_R)) = p_1(p_2(\mathbf{x}_R)) \wedge p_1(p_2(\mathbf{x}_R)) = p_2(p_2(\mathbf{x}_R)) \\ \bot & otherwise \end{cases}$$

$$\phi_{P2}(r^{out}) = \begin{cases} (p_1(p_1(\mathbf{x}_R)), [(p_1(p_2(\mathbf{x}_R)), p_2(p_2(\mathbf{x}_R)))]_{(\bot\bot),f} & p_1(p_1(\mathbf{x}_R)) = p_1(p_2(\mathbf{x}_R)) \wedge p_1(p_2(\mathbf{x}_R)) = p_2(p_2(\mathbf{x}_R)) \\ \bot & otherwise \end{cases}$$

As the filter conditions in both programs are equal, we only need to prove, with the knowledge that:

$$(*) p_1(p_1(\mathbf{x}_R)) = p_1(p_2(\mathbf{x}_R)) \wedge p_1(p_2(\mathbf{x}_R)) = p_2(p_2(\mathbf{x}_R))$$

the following:

$$(p_1(p_1(\mathbf{x}_R)), (p_1(p_2(\mathbf{x}_R)), [p_2(p_2(\mathbf{x}_R))]_{\bot,g})) = (p_1(p_1(\mathbf{x}_R)), [(p_1(p_2(\mathbf{x}_R)), p_2(p_2(\mathbf{x}_R)))]_{(\bot\bot),f})$$

Which in its turn can be simplified to proving:

$$(p_1(p_2(\mathbf{x}_R)), [p_2(p_2(\mathbf{x}_R))]_{\bot,g}) = [(p_1(p_2(\mathbf{x}_R)), p_2(p_2(\mathbf{x}_R)))]_{(\bot\bot),f}$$

We write $X = p_1(p_1(\mathbf{x}_R)) = p_1(p_2(\mathbf{x}_R)) = p_2(p_2(\mathbf{x}_R))$, and simplify:

$$(X, [X]_{\bot,g}) = [(X, X)]_{(\bot,\bot),f}$$

The sloppy handling of *foldByKey* representative elements brought us to an equivalence formula which is not provable (as X ranges over all of the integer domain). But, under the $(*)$ assumption, we rewrite some of the representative elements using the fold by key operator:

$$\phi_{P1}(R_3) = \langle((p_1(p_1(\mathbf{x}_R)), p_1(p_2(\mathbf{x}_R))), p_2(p_2(\mathbf{x}_R)))\rangle_{\bot,g}$$

$$\phi_{P1}(r^{out}) = (p_1(p_1(\phi_{P1}(R_3))), (p_2(p_1(\phi_{P1}(R_3))), p_2(\phi_{P1}(R_3))))$$

$$\phi_{P2}(r^{out}) = \langle \mathbf{x}_R \rangle_{(\bot\bot),f}$$

Under the $(*)$ assumption, the following is translated to:

$$\phi_{P1}(R_3) = \langle((X, X), X)\rangle_{\bot,g}$$

$$\phi_{P1}(r^{out}) = (p_1(p_1(\langle((X, X), X)\rangle_{\bot,g})), (p_2(p_1(\langle((X, X), X)\rangle_{\bot,g})), p_2(\langle((X, X), X)\rangle_{\bot,g})))$$

$$\phi_{P2}(r^{out}) = \langle(X, (X, X))\rangle_{(\bot\bot),f}$$

According to the rule defined in lemma 11, all instances of the key variables in an RDD element in the fold functions f and g become constant. In our case, the key variable is also the value variable, which is X. Thus, $f_k == (X, X)$, $g_k == X$, and we get in both programs, that for each element of the RDD R satisfying $(*)$, an element of the form $(X, (X, X))$. ■

This example may be a bit contrived, but it shows several properties:

- It is possible to consider keys of variable length if mapped correctly, even though real-life code hardly contain such abnormalities.
- Under certain conditions, it is possible to compare aggregated values with an RDD's elements
- There is a great importance to keeping track of the key variables of the representative element, after any kind of operation applied on the RDD. This allows to fixate the values correctly.

6 Related Work

- [11] contains a formal discussion of a logic with aggregate operators and arithmetic functions over $\mathbb{Q}$, and an embedding of a commercial SQL fragment (and its common aggregate functions) to the aggregate logic.
- [14] describe an extension of Presburger arithmetic with boolean algebras of sets of uninterpreted elements, with capabilites to describe correlations between integer variables and set cardinalities, which is decidable. One of the applications of their theory is constraint database query evaluation.
- [13] is a non-recent work reviewing

7 Future work

- Sound and complete handling of unions, also in aggregates
- Defining abstractions for Spark programs.
- Check properties of Spark programs, not just equivalence/equivalence modulo a transform
- Sound verification of programs with conditions in the program body, not just in UDFs.
- Loops, both in program body and UDFs. This will allow verifying a wider range of programs:

  ▸ Example 7.1. Nice example with finding the median:

```
Algo 1:
val C = R.fold(0,count);
val median = ⊥;
for i = 0; i < C;
        val min = R.fold(⊥,min);
        val countMin = R.filter(x == min).fold(0,count);
        i = i + countMin;
        if (2*i >= C)
                median = min;
                break;
        R = R.filter(x != min);
```

```
Algo 2:
Replace min with max.
```

  These programs both compute the median, using different aggregate functions.
- Implementing the procedure using Cooper's algorithm for Presburger arithmetic.
- Adding ordering to our definition of RDDs?

SG: TODO remove future tense from the rest of the paper

1 Appendix 1: Typing rules for $SPARK$

Booleans

$$\frac{}{\rho \vdash \texttt{true}:Boolean} \qquad \frac{}{\rho \vdash \texttt{false}:Boolean}$$

Integers

$$\frac{}{\rho \vdash 0,1,\ldots:Integer}$$

Integer ops

$$\frac{\rho \vdash i:Integer, j:Integer, op \in \{+,-,*,\%\}}{\rho \vdash i \ op \ j:Integer} \qquad \frac{\rho \vdash i:Integer, j:Integer, op \in \{<,\leq,=,\geq,>\}}{\rho \vdash i \ op \ j:Boolean}$$

Boolean ops

$$\frac{\rho \vdash b:Boolean}{\rho \vdash !b:Boolean} \qquad \frac{\rho \vdash b_1:Boolean, b_2:Boolean, op \in \{\wedge,\vee\}}{\rho \vdash b_1 \ op \ b_2:Boolean}$$

Tuples

$$\frac{\rho \vdash e_1:\tau_1, e_2:\tau_2}{\rho \vdash (e_1,e_2):\tau_1 \times \tau_2} \qquad \frac{\rho \vdash e:\tau_1 \times \ldots \times \tau_n}{\rho \vdash p_i(e):\tau_i}$$

UDFs

$$\frac{\rho \vdash f:C_1 \times \cdots \times C_n \to (\tau \to \tau'), \vec{e}:C_1 \times \cdots \times C_n}{\rho \vdash f(\vec{e}):\tau \to \tau'} \qquad \frac{\rho \vdash f:\tau \to \tau', t:\tau}{\rho \vdash f(t):\tau'}$$

RDD

$$\frac{\rho \vdash r:RDD_\tau, f:\tau \to \tau'}{\rho \vdash \texttt{map}(f)(r):RDD_{\tau'}} \qquad \frac{\rho \vdash r:RDD_\tau, f:\tau \to Boolean}{\rho \vdash \texttt{filter}(f)(r):RDD_\tau}$$

$$\frac{\rho \vdash r:RDD_\tau, r':RDD_{\tau'}}{\rho \vdash \texttt{cartesian}(r,r'):RDD_{\tau \times \tau'}}$$

$$\frac{\rho \vdash r:RDD_\tau, f:\tau' \times \tau \to \tau, init:\tau'}{\rho \vdash \texttt{fold}(init,f)(r):\tau'} \qquad \frac{\rho \vdash r:RDD_{K \times V}, f:(V' \times V) \to V', init:V'}{\texttt{foldByKey}(init,f)(r):RDD_{K \times V'}}$$

Figure 13 Typing rules for $SPARK$

2 Appendix 2: Proof of the equivalence of the standard semantics and SR(P) semantics for $SPARK$

The proof follows by structural induction on the available operations in the $SPARK$ program P (The syntactic term E). We also assume w.l.o.g. the RDDs given as arguments to the operations are input RDDs.

- Input RDD : Suppose $E = r, r \in \vec{r^{in}}$.

$$\begin{aligned} SR(P)(\vec{r^{in}}) = \{\!\{\mathbf{x}_r[\mathbf{x}_r \mapsto x] \mid \mathbf{x}_r[\mathbf{x}_r \mapsto x] \neq \bot \wedge x \in r\}\!\} &= \{\!\{x \mid x \neq \bot x \in r\}\!\} \\ &= \{\!\{x; r(x) \mid x \in r\}\!\} \\ &= [\![P]\!](\vec{r^{in}}) \end{aligned}$$

- *map*: Suppose $E = \texttt{map}(f)(r)$, $r \in \vec{r^{in}}$. We have

$$\begin{aligned} SR(P)(\vec{r^{in}}) &= \{\!\{\Phi(P)[\mathbf{x}_r \mapsto x] \mid \Phi(P)[\mathbf{x}_r \mapsto x] \neq \bot \wedge x \in r\}\!\} \\ &= \{\!\{f(\mathbf{x}_r)[\mathbf{x}_r \mapsto x] \mid f[\mathbf{x}_r \mapsto x] \neq \bot \wedge x \in r\}\!\} \\ &= \{\!\{f(x) \mid f(x) \neq \bot \wedge x \in r\}\!\} \end{aligned}$$

  While in the original semantics:

$$[\![P]\!] = [\![\texttt{map}]\!](f)(r) = \pi_2(\{\!\{(x, f(x)); r(x) \mid x \in r\}\!\})$$

  Recall that $[\![\texttt{map}]\!]$ returns a bag, so by taking some y such that $y \in [\![\texttt{map}]\!](f)(r)$, we know how to calculate its multiplicity in the bag:

$$([\![\texttt{map}]\!](f)(r))(y) = \sum_{(x,f(x)) \in \{\!\{(x,f(x));r(x) \mid x \in r\}\!\} \wedge f(x) = y} \{\!\{(x, f(x)); r(x) \mid x \in r\}\!\}(x, f(x)) = \sum_{x \in r \wedge f(x) = y} r(x)$$

  which is the canonical representation of the bag defined by $SR(P)(\vec{r^{in}})$ - the multiplicity of $f(x)$ is equal to the sum of all possible preimages' multiplicities $r(x)$ in r.
- *filter*: Suppose $E = \texttt{filter}(f)(r)$, $r \in \vec{r^{in}}$.

$$\begin{aligned} SR(P)(r) &= \{\!\{\Phi(P)[\mathbf{x}_r \mapsto x] \mid \Phi(P)[\mathbf{x}_r \mapsto x] \neq \bot \wedge x \in r\}\!\} \\ &= \{\!\{ite(f(\mathbf{x}_r) = tt, \mathbf{x}_r, \bot)[\mathbf{x}_r \mapsto x] \mid ite(f(\mathbf{x}_r) = tt, \mathbf{x}_r, \bot)[\mathbf{x}_r \mapsto x] \neq \bot \wedge x \in r\}\!\} \\ &= \{\!\{ite(f(x) = tt, x, \bot) \mid ite(f(x) = tt, x, \bot) \neq \bot \wedge x \in r\}\!\} \\ &= \{\!\{x \mid f(x) = tt \wedge x \neq \bot \wedge x \in r\}\!\} \\ &= \{\!\{x \mid f(x) = tt \wedge x \in r\}\!\} \end{aligned}$$

  While:

$$[\![\texttt{filter}]\!](f)(r) = r \restriction_{\{x \mid f(x)\}} = \sigma_{x \in \{x \mid f(x)\}}(r) = \{\!\{x; r(x) \mid x \in r \wedge x \in \{x \mid f(x)\}\}\!\}$$

  And the equality of the bags follows.
- *cartesian*: Suppose $E = \texttt{cartesian}(r_1, r_2)$, $r_1, r_2 \in \vec{r^{in}}$. We have:

$$\begin{aligned} SR(P)(r_1, r_2) &= \{\!\{\Phi(P)[\mathbf{x}_{r_1}^{(1)} \mapsto x_1, \mathbf{x}_{r_2}^{(2)} \mapsto x_2] \mid \Phi(P)[\mathbf{x}_{r_1}^{(1)} \mapsto x_1, \mathbf{x}_{r_2}^{(2)} \mapsto x_2] \neq \bot \wedge x_1 \in r_1 \wedge x_2 \in r_2\}\!\} \\ &= \{\!\{(\mathbf{x}_{r_1}^{(1)}, \mathbf{x}_{r_1}^{(2)})[\mathbf{x}_{r_1}^{(1)} \mapsto x_1, \mathbf{x}_{r_2}^{(2)} \mapsto x_2] \mid (\mathbf{x}_{r_1}^{(1)}, \mathbf{x}_{r_1}^{(2)})[\mathbf{x}_{r_1}^{(1)} \mapsto x_1, \mathbf{x}_{r_2}^{(2)} \mapsto x_2] \neq \bot \wedge x_1 \in r_1 \wedge x_2 \in \\ &= \{\!\{(x_1, x_2) \mid (x_1, x_2) \neq \bot \wedge x_1 \in r_1 \wedge x_2 \in r_2\}\!\} \\ &= \{\!\{(x_1, x_2) \mid x_1 \in r_1 \wedge x_2 \in r_2\}\!\} \end{aligned}$$

  In the standard semantics, we have:

$$[\![\texttt{cartesian}]\!](r_1, r_2) = r_1 \times r_2 = \{\!\{(x_1, x_2)\,;\, r_1(x_1) \cdot r_2(x_2) \mid x_1 \in r_1 \wedge x_2 \in r_2\}\!\}$$

  And the equality is straightforward.

- *fold*: Suppose $E = \texttt{fold}(e,f)(r), r \in \vec{r^{in}}$. We have: $SR(P)(\vec{r^{in}}) =^{def} [\mathbf{x}_r]_{e,f} =^{def} [\![\texttt{fold}]\!](e,f)(r)$. Equivalence by definition.
- *foldByKey*: Suppose $E = \texttt{foldByKey}(e,f)(r)$. We have:

$$SR(P)(\vec{r^{in}}) \quad =^{def} \{\!\{\langle \mathbf{x}_r \rangle_{e,f} \mid \mathbf{x}_r \in r\}\!\} = \{\!\{(p_1(\mathbf{x}_r), [p_2(\mathbf{x}_r)]_{e,f} \mid \mathbf{x}_r \in r\}\!\}$$

And by definition, the set of free variables for $SR(P)(\vec{r^{in}})$ is $p_1(\mathbf{x}_r)$. By proving: $[p_2(\mathbf{x}_r)]_{e,f} = [\![\texttt{fold}]\!](e,f)(\pi_2(r \restriction_{(p_1(\mathbf{x}_r),\_)}))$ we get $SR(P)(r) = [\![\texttt{foldByKey}]\!](e,f)(r)$, as required. Indeed, $r \restriction_{(p_1(\mathbf{x}_r),\_)} = \{\!\{x; r(x) \mid p_1(x) = p_1(\mathbf{x}_r)\}\!\}$ and $(\pi_2(r \restriction_{(p_1(\mathbf{x}_r),\_)})) = \{\!\{p_2(x); \sum_{p_2(y)=p_2(x) \wedge p_1(y)=p_1(\mathbf{x}_r)} r(y) \mid p_1(x) = p_1(\mathbf{x}_r)\}\!\}$.

Need to consider non-input RDD expressions

3 Appendix 3: Examples with Cartesian Product and Join

We shall define the *join* operator based on the *cartesian*, *filter* and *map* operators, and then proceed with examples proving properties of the join.

▸ Example 3.1 (Natural Join). An implementation of the *natural join* with existing operations.

	P1(R: $RDD_{K\times V}$, R': $RDD_{K\times V'}$):	P2(R: $RDD_{K\times V}$, R': $RDD_{K\times V'}$):
1	`return join`(R, R')	C = `cartesian`(R, R')
2		J = `filter`$(\lambda x.p_1(p_1(x)) == p_1(p_2(x)))(C)$
3		`return map`$(\lambda x.(p_1(p_1(x)), (p_2(p_1(x)), p_2(p_2(x)))))(J)$

Let us mark the representative elements of the variables of $P2$:
Inputs: $\mathbf{x}_R, \mathbf{x}_{R'}$, shorthand for $(p_1(\mathbf{x}_R), p_2(\mathbf{x}_R)), (p_1(\mathbf{x}_{R'}), p_2(\mathbf{x}_{R'}))$

$$\phi_{P2}(C) = (\mathbf{x}_R, \mathbf{x}_{R'})$$

$$\phi_{P2}(J) = \begin{cases} \phi_{P2}(C) & p_1(p_1(\phi_{P2}(C))) = p_1(p_2(\phi_{P2}(C))) \\ \bot & otherwise \end{cases} = \begin{cases} (\mathbf{x}_R, \mathbf{x}_{R'}) & p_1(\mathbf{x}_R) = p_1(\mathbf{x}_{R'}) \\ \bot & otherwise \end{cases}$$

$$\Phi(P2) = \begin{cases} (p_1(\mathbf{x}_R), (p_2(\mathbf{x}_R), p_2(\mathbf{x}_{R'})) & p_1(\mathbf{x}_R) = p_1(\mathbf{x}_{R'}) \\ \bot & otherwise \end{cases}$$

We see that the representative element of the returned RDD expression is describing exactly the expected semantics of the natural join. We can proceed and use *join* in our examples as syntactic sugar for a combination of *cartesian*, *filter* and *map* in the method that fits the arity of the tuple types automatically, similarly to the example.
The representative element for the *join* of RDDs R: $RDD_{K\times V}$, R': $RDD_{K\times W}$ is:

$$\phi(R \bowtie R') = \begin{cases} (p_1(\varphi), (p_2(\varphi), p_2(\varphi')) & p_1(\varphi) = p_1(\varphi') \wedge \varphi = \phi(R), \varphi' = \phi(R') \\ \bot & otherwise \end{cases}$$

▸ Example 3.2 (*join* distributivity with *map*). This example shows a case where *join* is distributive with respect to *map*.

	P1(R_0: $RDD_{\text{Int}\times\text{Int}}$, R_1: $RDD_{\text{Int}\times\text{Int}}$):	P2(R_0: $RDD_{\text{Int}\times\text{Int}}$, R_1: $RDD_{\text{Int}\times\text{Int}}$):
1	`return map`$(\lambda x.2 * x)$(`join`(R_0, R_1))	S_0 = `map`$(\lambda x.2 * x)(R_0)$
2		S_1 = `map`$(\lambda x.2 * x)(R_1)$
3		`return join`(S_0, S_1)

For $P1$:

$$\Phi(P1) = \begin{cases} 2 * ((p_1(\mathbf{x}_{R_0}), (p_2(\mathbf{x}_{R_0}), p_2(\mathbf{x}_{R_1}))) & p_1(\mathbf{x}_{R_0}) = p_1(\mathbf{x}_{R_1}) \\ \bot & otherwise \end{cases}$$

For $P2$:

$$\phi_{P2}(S_0) = 2 * \mathbf{x}_{R_0}$$
$$\phi_{P2}(S_1) = 2 * \mathbf{x}_{R_1}$$

$$\Phi(P2) = \begin{cases} (p_1(\phi_{P2}(S_0)), (p_2(\phi_{P2}(S_0)), p_2(\phi_{P2}(S_1))) & p_1(\phi_{P2}(S_0)) = p_1(\phi_{P2}(S_1)) \\ \bot & otherwise \end{cases}$$

$$= \begin{cases} (p_1(2 * \mathbf{x}_{R_0}), (p_2(2 * \mathbf{x}_{R_0}), p_2(2 * \mathbf{x}_{R_1})) & p_1(2 * \mathbf{x}_{R_0}) = p_1(2 * \mathbf{x}_{R_1}) \\ \bot & otherwise \end{cases}$$

The equivalence formula $\forall \mathbf{x}_{R_0}, \mathbf{x}_{R_1}.\Phi(P1) = \Phi(P2)$ can be proven in the Presburger arithmetic using the schemes defined in proposition 4.3. ■

▸ Example 3.3 (A counterexample - no *join* distributivity with *map* when key mappings do not agree). This example shows when *join* is not distributive with respect to *map*.

	P1(R_0: $RDD_{\texttt{Int}\times\texttt{Int}}$, R_1: $RDD_{\texttt{Int}\times\texttt{Int}}$):	P2(R_0: $RDD_{\texttt{Int}\times\texttt{Int}}$, R_1: $RDD_{\texttt{Int}\times\texttt{Int}}$):
1	`return map`$(\lambda(x,y).(1,y))$(`join`(R_0, R_1))	S_0 = `map`$(\lambda(x,y).(1,y))(R_0)$
2		S_1 = `map`$(\lambda(x,y).(1,y))(R_1)$
3		`return join`(S_0, S_1)

Fpr $P1$:

$$\Phi(P1) = \begin{cases} (1, (p_2(\mathbf{x}_{R_0}), p_2(\mathbf{x}_{R_1}))) & p_1(\mathbf{x}_{R_0}) = p_1(\mathbf{x}_{R_1}) \\ \bot & otherwise \end{cases}$$

For $P2$:

$$\phi_{P2}(S_0) = (1, p_2(\mathbf{x}_{R_0}))$$
$$\phi_{P2}(S_1) = (1, p_2(\mathbf{x}_{R_1}))$$
$$\Phi(P2) = \begin{cases} (p_1(\phi_{P2}(S_0)), (p_2(\phi_{P2}(S_0)), p_2(\phi_{P2}(S_1))) & p_1(\phi_{P2}(S_0)) = p_1(\phi_{P2}(S_1)) \\ \bot & otherwise \end{cases}$$
$$= \begin{cases} (1, (p_2(\mathbf{x}_{R_0}), p_2(\mathbf{x}_{R_1}))) & 1 = 1 \\ \bot & otherwise \end{cases}$$

The two program terms are equivalent if and only if:

$$\forall \mathbf{x}_{R_0}, \mathbf{x}_{R_1}.p_1(\mathbf{x}_{R_0}) = p_1(\mathbf{x}_{R_1})$$

Or in other words, if the keys of the two input RDDs always agree. This is of course wrong, so there is no equivalency in this case.

▸ **Corollary 12.** *Let R, R' be RDDs with a pair type, the first type in the pair being K and second type being τ_1, τ_2 respectively. Let f be a function $f : K \times \tau_1 \to K' \times \sigma_1$, and g be a function $g : K \times \tau_2 \to K' \times \sigma_2$. If f and g agree on K (namely, $\forall x, y.p_1(x) = p_1(y) \iff p_1(f(x)) = p_1(g(y))$), then the following two programs are equivalent:*

$$(1) \quad \texttt{join}(\texttt{map}(f)(R), \texttt{map}(g)(R'))$$
$$(2) \quad \texttt{map}(\mathcal{F})(\texttt{join}(R, R'))$$

where $\mathcal{F} = \lambda(k, (v, w)).p_1(f((k, v))), (p_2(f((k, v))), p_2(g((k, w))))$

Proof. We know that $\forall x, y.p_1(x) = p_1(y) \iff p_1(f(x)) = p_1(g(y))$, and proceed to compare representative elements.

$$\phi_{\texttt{map}(f)(R)} = f(\mathbf{x}_R)$$
$$\phi_{\texttt{map}(g)(R')} = g(\mathbf{x}_{R'})$$
$$\phi_{\texttt{join}(\texttt{map}(f)(R), \texttt{map}(g)(R'))} = \begin{cases} (p_1(f(\mathbf{x}_R)), (p_2(f(\mathbf{x}_R)), p_2(g(\mathbf{x}_{R'})))) & p_1(f(\mathbf{x}_R)) = p_1(g(\mathbf{x}_{R'})) \\ \bot & otherwise \end{cases}$$
$$\phi_{\texttt{join}(R,R')} = \begin{cases} (p_1(\mathbf{x}_R), (p_2(\mathbf{x}_R), p_2(\mathbf{x}_{R'})) & p_1(\mathbf{x}_R)) = p_1(\mathbf{x}_{R'}) \\ \bot & otherwise \end{cases}$$

$$\phi_{\texttt{map}(\mathcal{F})(\texttt{join}(R,R'))} = \begin{cases} \mathcal{F}((p_1(\mathbf{x}_R),(p_2(\mathbf{x}_R),p_2(\mathbf{x}_{R'}))) & p_1(\mathbf{x}_R)) = p_1(\mathbf{x}_{R'}) \\ \bot & otherwise \end{cases}$$
$$= \begin{cases} (p_1(f(\mathbf{x}_R)),(p_2(f(\mathbf{x}_R)),p_2(g(\mathbf{x}_{R'})))) & p_1(\mathbf{x}_R)) = p_1(\mathbf{x}_{R'}) \\ \bot & otherwise \end{cases}$$

All that is left to prove is the equivalence of the conditions:

$$\forall \mathbf{x}_R, \mathbf{x}_{R'}. p_1(f(\mathbf{x}_R)) = p_1(g(\mathbf{x}_{R'})) \iff p_1(\mathbf{x}_R)) = p_1(\mathbf{x}_{R'})$$

Which is proved directly from the known property of f, g of agreeing on K. ◂

Self products When handling self-joins or cartesian product on the same RDD, it is crucial to rename the free variables of the representative elements. For example:

	P1(R: $RDD_{\texttt{Int}}$):	P2(R: $RDD_{\texttt{Int}}$):
1	`return cartesian`(R,R)	`return map`$(\lambda x.(x,x))(R)$

The two programs are not equivalent due to different free variable sets. The program term of $P1$ is $(\phi_{P1}(R), \phi_{P1}(R))$, while the program term of $P2$ is: $(\mathbf{x}_R, \mathbf{x}_R)$. If ϕ did not rename variables, we would get: $\phi_{P1} = (\mathbf{x}_R, \mathbf{x}_R) = \phi_{P2}$ which is a mistake. Therefore ϕ is designed to generate *fresh* variables for RDDs.

▸ Example 3.4 (Self joins). Given an RDD of integers, we want all pairs of elements in the cartesian product whose sum is greater than 100. We can filter out all pairs where both elements are not larger than 50.

	P1(R: $RDD_{\texttt{Int}}$):	P2(R: $RDD_{\texttt{Int}}$):
1	C = `cartesian`(R,R)	C = `cartesian`(R,R)
2	`return filter`$(\lambda x,y.x+y>100)(C)$	C' = `filter`$(\lambda x,y.x>50 \vee y>50)(C)$
3		`return filter`$(\lambda x,y.x+y>100)(C')$

The representative elements of $P1$ (note the automatic renaming of variables for R):

$$\mathbf{x}_C = (\phi_{P1}(R), \phi_{P1}(R)) = (\mathbf{x}_R^{(1)}, \mathbf{x}_R^{(2)})$$
$$\Phi(P1) = \begin{cases} (\mathbf{x}_R^{(1)}, \mathbf{x}_R^{(2)}) & \mathbf{x}_R^{(1)} + \mathbf{x}_R^{(2)} > 100 \\ \bot & otherwise \end{cases}$$

The representative elements of $P2$:

$$\phi_{P2}(C) = (\mathbf{x}_R^{(1)}, \mathbf{x}_R^{(2)})$$
$$\phi_{P2}(C') = \begin{cases} (\mathbf{x}_R^{(1)}, \mathbf{x}_R^{(2)}) & \mathbf{x}_R^{(1)} > 50 \vee \mathbf{x}_R^{(2)} > 50 \\ \bot & otherwise \end{cases}$$
$$\Phi(P2) = \begin{cases} \begin{cases} (\mathbf{x}_R^{(1)}, \mathbf{x}_R^{(2)}) & \mathbf{x}_R^{(1)} > 50 \vee \mathbf{x}_R^{(2)} > 50 \\ \bot & otherwise \end{cases} & \mathbf{x}_R^{(1)} + \mathbf{x}_R^{(2)} > 100 \\ \bot & otherwise \end{cases}$$

We can show equivalency if we prove:

$$((\mathbf{x}_R^{(1)} > 50 \vee \mathbf{x}_R^{(2)} > 50) \wedge \mathbf{x}_R^{(1)} + \mathbf{x}_R^{(2)} > 100) \iff \mathbf{x}_R^{(1)} + \mathbf{x}_R^{(2)} > 100$$

A proof is implemented in Figure 14. ■

```
integer_ qelim
<<forall x,y.
((x>50 \/ y>50) /\ x+y>100)<=>(x+y>100)>>;;
- :  fol formula = <<true>>
```

Figure 14 Output of Cooper ML implementation proving this example

4 Appendix 4: Proof of soundness and completeness of inductive method for aggregates

▸ **Lemma 13.** *Let* $R_0 \in RDD_{\sigma_0}, R_1 \in RDD_{\sigma_1}$*, and denote their representative elements* φ_0, φ_1 *respectively. We assume* φ_0, φ_1 *were composed from* map *,* filter *and cartesian product (without self products). Let there be two* fold *functions* $f_0 : \xi_0 \times \sigma_0 \to \xi_0, f_1 : \xi_1 \times \sigma_1 \to \xi_1$*, two* initial *values* $init_0 : \xi_0, init_1 : \xi_1$*, and two functions* $g : \xi_0 \to \xi, g' : \xi_1 \to \xi$*. We have: if*

$$FV(\varphi_0) \simeq FV(\varphi_1), \text{ denoted } FV \tag{1}$$

$$g(init_0) = g'(init_1) \tag{2}$$

$$\forall \vec{v}, A_{\varphi_0} : \xi_0, A_{\varphi_1} : \xi_1 . g(A_{\varphi_0}) = g'(A_{\varphi_1}) \implies g(f_0(A_{\varphi_0}, \varphi_0[FV \mapsto \vec{v}])) = g'(f_1(A_{\varphi_1}, \varphi_1[FV \mapsto \vec{v}])) \tag{3}$$

then $g([\varphi_0]_{init_0,f_0}) = g'([\varphi_1]_{init_1,f_1})$

Proof. First we recall the semantics of the `fold` operation on some RDD R, which is a bag. We choose an arbitrary element $a \in R$ and apply the fold function recursively on a and on R with a single instance of a removed. We then write a sequence of elements in the order they are chosen by `fold`: $\langle a_1, \ldots, a_n \rangle$, where n is the sum of all multiplicities in the bag R. We also know that a requirement of aggregating operations' UDFs is that they are *commutative* and *associative*, so the order of elements chosen does not change the final result. We also extend f_i to $\xi_i \times (\sigma_i \cup \{\bot\})$ by setting $f_i(A, \bot) = A$ ($\bot$ is defined to behave as the neutral element for f_i). To prove $g([\varphi_0]_{init_0,f_0}) = g'([\varphi_1]_{init_1,f_1})$, it is necessary to prove that

$$g([\![\texttt{fold}]\!](f_0, init_0)(R_0)) = g'([\![\texttt{fold}]\!](f_1, init_1)(R_1))$$

We set $A_{\varphi_j,0} = init_j$ for $j \in \{0, 1\}$. Each element of R_0, R_1 is expressible by providing a concrete valuation to the free variables of φ_0, φ_1, namely the vector $\vec{v}$. We choose an arbitrary sequence of valuations to $\vec{v}$, denoted $\langle \vec{a}_1, \ldots, \vec{a}_n \rangle$, and plug them into the *fold* operation for both R_0, R_1. The result is 2 sequences of *intermediate values* $\langle A_{\varphi_0,1}, \ldots, A_{\varphi_0,n} \rangle$ and $\langle A_{\varphi_1,1}, \ldots, A_{\varphi_1,n} \rangle$. We have that $A_{\varphi_j,i} = f_j(A_{\varphi_j,i-1}, \varphi_j[FV \mapsto \vec{a}_i])$ for $j \in \{0, 1\}$, from the semantics of `fold`. Our goal is to show $g(A_{\varphi_0,n}) = g'(A_{\varphi_1,n})$ for all n. We prove the equality by induction on the *size* of the sequence of possible valuations of $\vec{v}$, denoted n. In each step i, we show $g(A_{\varphi_0,i}) = g'(A_{\varphi_1,i})$.

Case $n = 0$: $R_0 = R_1 = \bot$, so $[\![\texttt{fold}]\!](f_0, init_0)(R_0) = init_0$ and $[\![\texttt{fold}]\!](f_1, init_1)(R_1) = init_1$. From Equation (1), $g(init_0) = g'(init_1)$, as required.

Case $n = i$, assuming correct for $n \le i - 1$: By assumption, we know that the sequence of intermediate values up to $i - 1$ is equal up to application of g, g', and specifically $g(A_{\varphi_0,i-1}) = g'(A_{\varphi_1,i-1})$. We are given the i'th concrete valuation of $\vec{v}$, denoted $\vec{a}_i$. We need to show $A_{\varphi_0,i} = A_{\varphi_1,i}$, so we use the formula for calculating the next intermediate value:

$$\begin{aligned} A_{\varphi_0,i} &= f_0(A_{\varphi_0,i-1}, \varphi_0[FV \mapsto \vec{a}_i]) \\ A_{\varphi_1,i} &= f_1(A_{\varphi_1,i-1}, \varphi_1[FV \mapsto \vec{a}_i]) \end{aligned}$$

We use Equation (2), plugging in $\vec{v} = \vec{a}_i$, $A_{\varphi_0} = A_{\varphi_0,i-1}$, and $A_{\varphi_1} = A_{\varphi_1,i-1}$. By the induction assumption, $g(A_{\varphi_0,i-1}) = g'(A_{\varphi_1,i-1})$, therefore $g(A_{\varphi_0}) = g'(A_{\varphi_1})$, so Equation (2) yields $g(f_0(A_{\varphi_0}, \varphi_0[FV \mapsto \vec{a}_i])) = g'(f_1(A_{\varphi_1}, \varphi_1[FV \mapsto \vec{a}_i]))$. By substituting back A_{φ_j} and the formula for the next intermediate value, we get: $g(A_{\varphi_0,i}) = g'(A_{\varphi_1,i})$ as required. ◂

5 Appendix 5: Examples with aggregation (*fold*)

We present several examples, showing different combinations of *fold*, various UDFs, and relational operators.

▸ Example 5.1 (Double counting). Below is a basic example of an application of lemma 3.

Let: $f = \lambda A,(a,b).A+b$

	P1(R: RDD_τ):	P2(R: RDD_τ):
1	$R' = \texttt{map}(\lambda x.(x,1))(R)$	$R' = \texttt{map}(\lambda x.(x,2))(R)$
2	$\texttt{return } 2 * \texttt{fold}(0,f)(R')$	$\texttt{return fold}(0,f)(R')$

After calculating representative elements for R' in both programs (which is straightforward), for equivalence we need to prove $2 * [(\mathbf{x}_r,1)]_{0,f} = [(\mathbf{x}_r,2)]_{0,f}$. We apply Lemma 3. We set $g(x) = id, g'(x) = (\lambda x.2 * x)$, both *init* values to 0, and check the two conditions:

$$0 = g(0) = 2 * 0 \tag{1}$$

$$\forall x, A, A'.g(A) = A' \implies \tag{2}$$
$$g(f(A,(x,1))) = g(A+1) = 2 * A + 2 = g(A) + 2 = A' + 2 = f(A',(x,2))$$

Which is what had to be proven. ■

▸ Example 5.2 (Sum with/without zeroes). The next example deals with the ability to handle neutral elements of summation.

Let: $sum = \lambda A, x.A + x$

	P1(R: $RDD_{\texttt{Int}}$):	P2(R: $RDD_{\texttt{Int}}$):
1	$R' = \texttt{filter}(\lambda x.x > 0)(R)$	$R' = \texttt{filter}(\lambda x.x > -1)(R)$
2	$\texttt{return fold}(0, sum)(R')$	$\texttt{return fold}(0, sum)(R')$

The program terms:

$$\Phi(P1) = \left[\begin{cases}\mathbf{x}_R & \mathbf{x}_R > 0\\ \bot & otherwise\end{cases}\right]_{0,sum}; \Phi(P2) = \left[\begin{cases}\mathbf{x}_R & \mathbf{x}_R > -1\\ \bot & otherwise\end{cases}\right]_{0,sum}$$

For *init* case, equivalency is obvious ($0 = 0$). For the induction step we know that $\bot$ valued-elements do not change the intermediate value. In our case, knowing that the operation is summation, and for brevity of writing, we replace all $\bot$ instances in the representative element with the neutral element of the given *fold* UDF, which is 0. We assume $A = A'$ and need to prove:

$$A + \begin{cases}\mathbf{x}_R & \mathbf{x}_R > 0\\ 0 & otherwise\end{cases} = A' + \begin{cases}\mathbf{x}_R & \mathbf{x}_R > -1\\ 0 & otherwise\end{cases}$$

Which boils down to the following comparison according to entire the 2×2 matrix of possible cases:

$$(\mathbf{x}_R > 0 \wedge \mathbf{x}_R > -1 \wedge \mathbf{x}_R = \mathbf{x}_R) \vee (\mathbf{x}_R <= 0 \wedge \mathbf{x}_R > -1 \wedge \mathbf{x}_R = 0) \vee$$
$$(\mathbf{x}_R <= 0 \wedge \mathbf{x}_R <= -1 \wedge 0 = 0) \vee (\mathbf{x}_R > 0 \wedge \mathbf{x}_R <= -1 \wedge \mathbf{x}_R = 0)$$

A proof using an open-source implementation of Cooper's algorithm in OCaml [10, 2] is given in Figure 15. ■

```
integer_ qelim
<<forall x.
((x > 0 /\  x > -1) \/ ((x <= 0 /\  x > -1) /\  x = 0) \/ (x <= 0 /\  x <=
-1) \/ ((x > 0 /\  x <= -1) /\  x = 0))>>;;
- :  fol formula = <<true>>
```

Figure 15 Output of Cooper ML implementation proving this example

▸ Example 5.3 (Divisibility by 9). Here we show two programs that get an RDD of integers, which is interpreted as a number. Every element in the RDD is a digit. By assuming that the iteration over the elements is in fixed order, the RDD indeed represents a unique number.

Let: $makeNumber = \lambda N, x.N * 10 + x, sum = \lambda S, x.S + x$

	P1(R: RDD_{Int}):	P2(R: RDD_{Int}):
1	v = `fold`$(0, makeNumber)(R)$	v = `fold`$(0, sum)(R)$
2	`return` $v\%9$	`return` $v\%9$

The program terms are:

$$\Phi(P1) = [\mathbf{x}_R]_{0,makeNumber}\%9$$
$$\Phi(P2) = [\mathbf{x}_R]_{0,sum}\%9$$

By applying lemma 3, taking $g = g' = \lambda x.x\%9$, we need to prove:

$$\forall x.A, A'.A\%9 = A'\%9 \implies (A * 10 + x)\%9 = (A' + x)\%9$$

Which is provable using Cooper's algorithm (see Figure 16). ■

```
integer_ qelim
<<forall x,a,b.
exists y,z.
((a-9*y = b-9*z)/\((a-9*y)<9)/\((a-9*y)>=0)) ==>
exists u,w.
(((10 * a + x)-9*u = (b+x)-9*w)/\(((b+x)-9*w) < 9)/\(((b+x)-9*w) >= 0))>>;;
- :  fol formula = <<true>>
```

Figure 16 Output of Cooper ML implementation proving this example

6 Appendix 6: Induction on self cartesian products aggregate expressions

As mentioned previously, self cartesian products, or more precisely, programs where a cartesian products has a non-empty intersection of the sets input RDD variables appearing in each of the tuple elements, are excluded. The following example shows why this exclusion was made:

▸ Example 6.1 (Aggregate on self join/cartesian - failure of lemma 3). We want to calculate $2n^2$ where n is the number of ones (1's) in an RDD.

	P1($R : RDD_{\texttt{Int}}$):	P2($R : RDD_{\texttt{Int}}$):
1	$O = \texttt{filter}(\lambda x.x = 1)(R)$	$O = \texttt{filter}(\lambda x.x = 1)(R)$
2	$C = \texttt{cartesian}(O, O)$	$O2 = \texttt{map}(\lambda x.2 * x)(O)$
3	$C' = \texttt{map}(\lambda(x, y).x + y)(C)$	$C2 = \texttt{cartesian}(O2, O2)$
4	$\texttt{return fold}(0, \lambda A, x.A + x)(C')$	$\texttt{return fold}(0, \lambda A, (x, y).A + y)(C2)$

For $P1, P2$:

$$\phi(O) = \begin{cases} \mathbf{x}_R & \mathbf{x}_R = 1 \\ \bot & otherwise \end{cases}$$

For $P1$:

$$\phi_{P1}(C) = (\phi_{P1}(O), \phi_{P1}(O))$$

$$\phi_{P1}(C') = \phi_{P1}(O) + \phi_{P1}(O) = \begin{cases} 1 & \mathbf{x}_R^{(1)} = 1 \\ \bot & otherwise \end{cases} + \begin{cases} 1 & \mathbf{x}_R^{(2)} = 1 \\ \bot & otherwise \end{cases}$$

$$\begin{aligned} \Phi(P1) &= [\phi_{P1}(C')]_{0,\lambda A,x.A+x} \\ &= \left[\begin{cases} 1 & \mathbf{x}_R^{(1)} = 1 \\ \bot & otherwise \end{cases} + \begin{cases} 1 & \mathbf{x}_R^{(2)} = 1 \\ \bot & otherwise \end{cases} \right]_{0,\lambda A,x.A+x} \end{aligned}$$

For $P2$:

$$\phi_{P2}(O2) = \begin{cases} 2 * \mathbf{x}_R^{(1)} & \mathbf{x}_R^{(1)} = 1 \\ \bot & otherwise \end{cases}$$

$$\phi_{P2}(C2) = (\phi_{P2}(O2), \phi_{P2}(O2))$$

$$\begin{aligned} \Phi(P2) &= [\phi_{P2}(C2)]_{0,\lambda A,(x,y).A+y} \\ &= \left[\left(\begin{cases} 2 & \mathbf{x}_R^{(1)} = 1 \\ \bot & otherwise \end{cases}, \begin{cases} 2 & \mathbf{x}_R^{(2)} = 1 \\ \bot & otherwise \end{cases} \right) \right]_{0,\lambda A,(x,y).A+y} \end{aligned}$$

Proceeding with the application of lemma 3: Let there be $x, y.\mathbf{x}_R^{(1)} = x, \mathbf{x}_R^{(2)} = y$ and intermediate values A, A', such that $A = A'$. Need to prove:

$$A + \begin{cases} 1 & x = 1 \\ \bot & otherwise \end{cases} + \begin{cases} 1 & y = 1 \\ \bot & otherwise \end{cases} = A' + \begin{cases} 2 & y = 1 \\ \bot & otherwise \end{cases}$$

If $x = 1, y = 1$, we get: $A + 1 + 1 = A' + 2$ which is true as induction assumption gives $A = A'$. If $x \neq 1$, then the fold function receives a $\bot$ and regards it as a neutral element, resulting in $A + 1 = A' + 2$, contradiction.

Handling self products The solution to the self-product problem is *multiple step inductions*. That is, we apply the induction step not on one element from the sequence, but on two or more elements. Furthermore, the elements are not chosen arbitrarily. We take advantage of the well-definition of *fold* operator and choose at once a sequence of elements which all belong to the same *symmetry class* in the complex RDD cartesian product with at least one variable appearing in both elements of the product. We start with illustrating the symmetry class for various RDDs:

1. $R \times R$ - symmetry classes are the singletons $\{(x,x)\}$ $(x \in R)$ (the diagonal of the product matrix), and $\{(x,y),(y,x)\}$ for $x,y \in R, x \neq y$.
2. $R \times R \times R$ - symmetry classes are the singletons $\{(x,x,x)\}$ $(x \in R)$, as well as the sets $\{(x,x,y),(x,y,x),(y,x,x),(x,y,y),(y,x,y),(y,y,x)\}$ for $x,y \in R, x \neq y$, and $\{(x,y,z),(x,z,y),(y,x,z),(y,z,x),(z,x,y),(z,y,x)\}$ for $x,y,z \in R, x \neq y \neq z$.
3. $R \times R \times R'$ - symmetry classes are similar to the first case: $\{(x,x,z)\}$ for $x \in R, z \in R'$ and $\{(x,y,z),(y,x,z)\}$ for $x,y \in R, x \neq y; z \in R'$
4. $R \times R' \times R \times R'$ - symmetry classes are similar to the previous case: $\{(x,y,x,y)\}$ for $x \in R, y \in R'$ and $\{(x,y,x',y'),(x,y',x',y),(x',y,x,y'),(x','y',x,y)\}$ for $x,x' \in R, x \neq x'; y,y' \in R', y \neq y'$.
5. $S = [\![\texttt{map}]\!](R \times R)(\lambda x,y.x+y)$ - symmetry classes are the same as in the first case. Even though we have a single RDD S, there are 2 symmetric valuations to it if the two elements of R are different, and 1 if they are equal.

In general, let there be a set of input RDDs $R_1,\ldots,R_n$ and an RDD $A = A_1 \times \cdot \times A_k$. For each $i \in \{1,\ldots,k\}$, let $\phi(A_i) = \varphi(R_{i_{j_1}},\ldots,R_{i_{j_i}})$, where $i_l \in \{1,\ldots,n\}$ for $l \in \{j_1,\ldots,j_i\}$. We denote the total number of appearances of an RDD R_i in A as c_i. The *symmetry family of R_i in A of size m* is the set of all symmetric partial valuations to A such that in the c_i appearances of R_i in A there are m unique variables. Each such symmetric partial valuation is a *symmetry class of R_i in A*. In general, the *symmetry family of R_i in A* is the union of the above for all $m \in \{1,\ldots,c_i\}$. For each symmetry family of size m we want to know what is the cardinality of each symmetry class, because this number will define the number of induction steps required for it. Illustratively, we have c_i cells that need to be filled with m unique values. We choose arbitrary m such values from the RDD R_i and build the symmetry class for it, denote the values $\{x_1,\ldots,x_m\}$. The number of times each time x_j value appears is denoted β_j, and the requirement is that $\sum_{j=1}^{m}\beta_j = c_i$. This is the same as choosing a partitioning of a set of size c_i to m non-empty sets with importance to the order of the sets in the partitioning. This is equal to $m! \cdot S(c_i,m)$ where $S(c_i,m)$ is the *Stirling set number* [] [2]. The closed formula for the size of a symmetry class for a symmetry family of size m of R in A is:

$$SCS(c_i,m) = \sum_{j=0}^{m}(-1)^j\binom{m}{j}(m-j)^{c_i}$$

Given some concrete valuation to all input RDDs, we can calculate for all $i \in \{1,\ldots,n\}$ the number m_i, which is the number of unique values chosen for the c_i instances of an RDD R_i. The number of induction steps required is the product of the symmetry class sizes for all RDDs:

$$\Pi_{i=1}^{n} SCS(c_i,m_i)$$

We now calculate it formally for the above examples:

[2] The *Stirling set number* $S(n,m)$ is the number of ways to partition a set of n elements to m non empty sets, and $m!$ is the number of possible orderings of the chosen sets.

1. For a valuation (x,x) we get $SCS(2,1) = \binom{1}{0}1^2 - \binom{1}{1}0^2 = 1$ as expected[3].
   For a valuation $(x,y), x \neq y$ we get $SCS(2,2) = \binom{2}{0}2^2 - \binom{2}{1}1^2 + \binom{2}{2}0^2 = 4-2 = 2$ as expected.
2. For a valuation (x,x,x) we get $SCS(3,1) = 1$ as expected.
   For a valuation $(x,x,y), x \neq y$ we get $SCS(3,2) = \binom{2}{0}2^3 - \binom{2}{1}1^3 + 0 = 8-2 = 6$ as expected.
   For a valuation $(x,y,z), x \neq y \neq z$ we get $SCS(3,3) = \binom{3}{0}3^3 - \binom{3}{1}2^3 + \binom{3}{2}1^3 - 0 = 27-24+3 = 6$, as expected.
3. For a valuation $(x_r, x_r, x_{r'})$ we get $SCS(2,1)SCS(1,1) = 1$ as expected.
   For a valuation $(x_r, y_r, x_{r'}), x_r \neq y_r$ we get $SCS(2,2)SCS(1,1) = 2$ as expected.
4. For a valuation $(x_r, x_{r'}, x_r, x_{r'})$ we get $SCS(2,1)SCS(2,1) = 1$ as expected.
   For a valuation $(x_r, x_{r'}, y_r, y_{r'}), x_r \neq y_r, x_{r'} \neq y_{r'}$ we get $SCS(2,2)SCS(2,2) = 4$ as expected.
5. Same as item 1.

Completing the proof of example 6.1 Reminder: we had $\mathbf{x}_R^{(1)} = x, \mathbf{x}_R^{(2)} = y$ and intermediate values A, A', such that $A = A'$, and needed to prove:

$$A + \begin{cases} 1 & x = 1 \\ \bot & otherwise \end{cases} + \begin{cases} 1 & y = 1 \\ \bot & otherwise \end{cases} = A' + \begin{cases} 2 & y = 1 \\ \bot & otherwise \end{cases}$$

We handle cases according to the number of choices to pick unique elements in the cartesian product (there are 2 such cases):

1. For $x = y$, we get that we need to prove:
$$A + \begin{cases} 1 & x = 1 \\ \bot & otherwise \end{cases} + \begin{cases} 1 & x = 1 \\ \bot & otherwise \end{cases} = A' + \begin{cases} 2 & x = 1 \\ \bot & otherwise \end{cases}$$
   which is immediate.
2. For $x \neq y$, we prove for 2 possible valuations: $\mathbf{x}_R^{(1)} = x, \mathbf{x}_R^{(2)} = y$ and $\mathbf{x}_R^{(1)} = y, \mathbf{x}_R^{(2)} = x$. We get that we need to prove:
$$\left(A + \begin{cases} 1 & x = 1 \\ \bot & otherwise \end{cases} + \begin{cases} 1 & y = 1 \\ \bot & otherwise \end{cases}\right) + \begin{cases} 1 & y = 1 \\ \bot & otherwise \end{cases} + \begin{cases} 1 & x = 1 \\ \bot & otherwise \end{cases} = \left(A' + \begin{cases} 2 & y = 1 \\ \bot & otherwise \end{cases}\right) + \begin{cases} 2 & x = 1 \\ \bot & otherwise \end{cases}$$
   which is indeed true, in all 3 possible cases where $x \neq y$: (1) $x = 1, y \neq 1$, or (2) $y = 1, x \neq 1$ or (3) $x \neq 1, y \neq 1, x \neq y$.

▸ **Lemma 14.** *Under the premises of lemma 3, where we allow self cartesian products too. For brevity, we denote the assignment of $\vec{v}$ to the free variables of φ_j as $\psi_j(\vec{v}) = \varphi_j[FV(\varphi_j) \mapsto \vec{v}]$. We denote $SF(n)$ as the symmetry family of both R_0, R_1 (true because $\mathcal{M}(R_0) = \mathcal{M}(R_1)$) of size n. In addition, $\sigma_i(\vec{v})$ will return the i'th permutation of the vector $\vec{v} \in SF(n)$ in its symmetry class, which is known to be of size $SCS(\mathcal{M}(R_0), n)$ - so $i \in \{1, \ldots, SCS(\mathcal{M}(R_0), n)\}$. We set $\sigma_1(\vec{v}) = \vec{v}$ for all $\vec{v}$. We have $g([\varphi_0]_{init_0, f_0}) = g'([\varphi_1]_{init_1, f_1})$ if:*

$$g(init_0) = g'(init_1) \tag{1}$$

[3] $\forall n. SCS(n,1) = 1$

$$\forall \vec{v}, A_{\varphi_0} : \xi_0, A_{\varphi_1} : \xi_1. \qquad g(A_{\varphi_0}) = g'(A_{\varphi_1}) \implies$$
$$\begin{cases} g(f_0(A_{\varphi_0}, \psi_0(\vec{v}))) = g'(f_1(A_{\varphi_1}, \psi_1(\vec{v}))) & \vec{v} \in SF(1) \\ g(f_0(f_0(A_{\varphi_0}, \psi_0(\vec{v}))), \psi_0(\sigma_2(\vec{v}))) = g'(f_1(f_1(A_{\varphi_1}, \psi_1(\vec{v}))), \psi_1(\sigma_2(\vec{v}))) & \vec{v} \in SF(2) \\ \qquad \ldots & \\ g(f_0(\ldots(f_0(A_{\varphi_0}), \psi_0(\sigma_1(\vec{v}))) \ldots), \sigma_{SCS(\mathcal{M}(R_0),n)}(\vec{v}))) \quad = \\ g'(f_1(\ldots(f_1(A_{\varphi_1}), \psi_1(\sigma_1(\vec{v}))) \ldots), \sigma_{SCS(\mathcal{M}(R_1),n)}(\vec{v}))) & \vec{v} \in SF(n) \end{cases} \tag{2}$$

The lemma above shows that the process described in this section allows sound and complete procedure for verifying PE, in the presence of self-cartesian-products.

References

1 Java. `http://java.net`. Accessed: 2016-07-19.

2 Ocaml implementation of cooper's algorithm. `https://www.cl.cam.ac.uk/~jrh13/atp/OCaml/cooper.ml`. Accessed: 2016-09-07.

3 Python. `https://www.python.org/`. Accessed: 2016-07-19.

4 Scala. `http://www.scala-lang.org`. Accessed: 2016-07-19.

5 Aaron R. Bradley and Zohar Manna. *The Calculus of Computation: Decision Procedures with Applications to Verification*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2007.

6 David C Cooper. Theorem proving in arithmetic without multiplication. *Machine Intelligence*, 1972.

7 Ron Cytron, Jeanne Ferrante, Barry K. Rosen, Mark N. Wegman, and F. Kenneth Zadeck. Efficiently computing static single assignment form and the control dependence graph. *ACM Trans. Program. Lang. Syst.*, 13(4):451–490, October 1991. URL: `http://doi.acm.org/10.1145/115372.115320`, `doi:10.1145/115372.115320`.

8 Jeffrey Dean and Sanjay Ghemawat. Mapreduce: Simplified data processing on large clusters. In *Proceedings of the 6th Conference on Symposium on Opearting Systems Design & Implementation - Volume 6*, OSDI'04, pages 10–10, Berkeley, CA, USA, 2004. USENIX Association. URL: `http://dl.acm.org/citation.cfm?id=1251254.1251264`.

9 Sanjay Ghemawat, Howard Gobioff, and Shun-Tak Leung. The google file system. In *Proceedings of the Nineteenth ACM Symposium on Operating Systems Principles*, SOSP '03, pages 29–43, New York, NY, USA, 2003. ACM. URL: `http://doi.acm.org/10.1145/945445.945450`, `doi:10.1145/945445.945450`.

10 John Harrison. *Handbook of Practical Logic and Automated Reasoning*. Cambridge University Press, New York, NY, USA, 1st edition, 2009.

11 Lauri Hella, Leonid Libkin, Juha Nurmonen, and Limsoon Wong. Logics with aggregate operators. *J. ACM*, 48(4):880–907, July 2001. URL: `http://doi.acm.org/10.1145/502090.502100`, `doi:10.1145/502090.502100`.

12 Ross Ihaka and Robert Gentleman. R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, 5(3):pp. 299–314, 1996.

13 Anthony C. Klug. Equivalence of relational algebra and relational calculus query languages having aggregate functions. *J. ACM*, 29(3):699–717, 1982. URL: `http://doi.acm.org/10.1145/322326.322332`, `doi:10.1145/322326.322332`.

14 Viktor Kuncak, Huu Hai Nguyen, and Martin C. Rinard. Deciding boolean algebra with presburger arithmetic. *J. Autom. Reasoning*, 36(3):213–239, 2006. `doi:10.1007/s10817-006-9042-1`.

15 Derek C. Oppen. A 222pn upper bound on the complexity of presburger arithmetic. *Journal of Computer and System Sciences*, 16(3):323 – 332, 1978. URL: http://www.sciencedirect.com/science/article/pii/0022000078900211, doi:http://dx.doi.org/10.1016/0022-0000(78)90021-1.

16 Matei Zaharia, Mosharaf Chowdhury, Tathagata Das, Ankur Dave, Justin Ma, Murphy McCauly, Michael J. Franklin, Scott Shenker, and Ion Stoica. Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing. In *Presented as part of the 9th USENIX Symposium on Networked Systems Design and Implementation (NSDI 12)*, pages 15–28, San Jose, CA, 2012. USENIX.