Hog Language Reference

Jason Halpern
Samuel Messing
Kurry Tran
Benjamin Rapaport
Paul Tylkin

March 17, 2012

Contents

Chapter 1

Introduction

As data sets have grown in size, so have the complexities of dealing with them. For instance, consider wanting to generate counts for all the words in *War and Peace* by means of distributed computation. Writing in Java and using Hadoop MapReduce (TM), a simple solution takes over 50 lines of code, as the programmer is required to specify intermediate objects not directly related to the desired computation, but required simply to get Hadoop to function properly. Our goal is to produce a language that can express the same computation in about 10 lines.

1.1 The MapReduce Framework

With the explosion in the size of datasets that companies have had to manage in recent years there are many new challenges that they face. Many companies and organizations have to handle the processing of datasets that are terabytes or even petabytes in size. The first challenge in this large-scale processing is how to make sense of all this data. More importantly, how can they process and manipulate the data in a time efficient and reliable manner. The second challenge is how they handle this across their distributed systems. Writing distributed, fault tolerant programs requires a high level of expertise and knowledge of parallel systems.

This was an obvious challenge to the company that has to process more data than any company on earth, Google. In response to this need, a group of engineers at Google developed their MapReduce framework in 2004. This high-level framework could be used for of a variety of tasks, including handling search queries, indexing crawled documents and processing logs. The software framework was developed to handle computations on massive datasets that are distributed across hundreds or even thousands of machines. The motivation behind MapReduce was to create a unified framework that abstracted away many of the low level details from programmers, so they would not have to be concerned with how the data is distributed, how the computation is parallelized and how all this is done in a fault tolerant manner. MapReduce provides fault tolerance in software rather than in hardware. MapReduce can handle both unstructured data (files) and structured data (databases), but is predominantly used with files.

The framework partitions the data across different machines, so that the computations are initially performed on smaller sets of data distributed across the cluster. Each cluster has a master node that is responsible for coordinating the efforts among the slave nodes. Each slave node sends periodic heartbeats to the master node so it can be aware of progress and failure. In the case of failure, the master node can reassign tasks to other nodes in the cluster. In conjunction with the underlying MapReduce framework created at Google, the company also had to build the distributed Google File System (GFS). This file system "allows programs to access files efficiently from any computer, so functions can be mapped everywhere." GFS was designed with the same goals as other distributed file systems, including "performance, scalability, reliability and availability." Another key aspect of the GFS design is fault tolerance and this is achieved by treating failures as normal and optimizing for "huge files that are mostly appended to and then read."

Within the framework, a programmer is responsible for writing both Map and Reduce functions. The map function is applied to all of the input data "in order to compute a set of intermediate key/value pairs." In the map step, the master node partitions the input data into smaller problems and distributes them across the worker nodes in the cluster. This step is applied in parallel to all of the input that has been partitioned across the cluster. Then, the reduce step is responsible for collecting all the processed data from the slave nodes and formatting the output. The reduce function is carried out over all the values that have the same key such that each key has a single value. which is the answer to the problem MapReduce is trying to solve. The output is done to files in the distributed file system.

The use of "a functional model with user-specified map and reduce operations allows (Google) to parallelize large computations easily and to use re-execution as the primary mechanism for fault tolerance." A programmer only has to specify the functions described above and the system handles the rest of the details. The following diagram illustrates the execution flow of a MapReduce program.

1.2 The Hog Language

Hog is a **data-oriented**, **high-level**, scripting language for creating MapReduce programs. Used alongside Hadoop, Hog enables users to efficiently carry out **distributed** computation. Hadoop MapReduce is an open-source framework for carrying out distributed computation, which is especially useful for working with large data sets. While it is possible to write code to carry out computations with Hadoop directly, the framework requires users to specify low-level details that are often irrelevant to their desired goal.

By building a scripting language on top of Hadoop, we aim to simplify the process. Built around a **simple** and highly **readable** syntax, Hog will let users focus on what computations they want done, and not how they want to do them. Hog takes care of all the low-level details required to run computations on Hadoops distributed network. All a user needs to do is tell Hog the location of their valid Hadoop instance, and Hog will do the rest.

TODO: about restrictions of Hog

1.2.1 Guiding Principles

The guiding principles of Hog are:

- Anyone can MapReduce
- Brevity over verbosity
- Preference simplicity over complexity

Chapter 2

Program Structure

2.1 Overall Structure

Every Hog program consists of a single source file with a .hog extension. This source file must contain three sections: `@Map`, and `@Reduce`, and `@Main` and can also include an optional `@Functions` section. These sections must be included in the following order:

```
@Functions
.
.
.
@Map <type signature>
.
.
.
@Reduce <type signature>
.
.
.
@Main
```

2.2 `@Functions`

At the top of every Hog program, the programmer has the option to define functions in a section called `@Functions`. Any function defined in this section can be called from any other section of the program, including `@Map`, and `@Reduce`, and can also be called from other functions defined in the `@Functions` section. The section containing the functions begins with the keyword `@Functions` on its own line, followed by the function definitions.

Function definitions have the form:

```
@Functions
return-type function-name(parameter-list) {
  exprlist
}
```

The return-type can be any valid Hog type. The rules regarding legal function-names are identical to those regarding legal variable identifiers. Each parameter in the parameter-list consists of a valid Hog type followed by the name of the parameter, which must also follow the naming rules for identifiers. Parameters in the parameter-list are separated by commas. The @Functions section ends when the next Hog section begins.

A complete example of an @Functions section:

```
@Functions
int min(int a, int b) {
  if (a < b) {
    return a
  }
  else {
    return b
  }
}

list<int> reverseList(list<int> oldList) {
  linst<int> newList()
  for (int i = oldList.len()-1; i >= 0; i--) {
    newList.append(oldList.get(i))
  }
  return newList
}
```

Function names can be overloaded as long as the function definitions have different signatures (i.e. parameter lists different in types and/or length). Additionally, user-defined functions can make reference to other user-defined functions.

2.3 @Map

The map function in a MapReduce program takes as input key-value pairs, performs the appropriate calculations and procedures, and emits intermediate key-value pairs as output. Any given input pair may map to zero, one, or multiple output pairs. The `@map` section defines the code for the map function.

The `@map` header must be followed by the signature of the map function, and then the body of the map function as follows:

```
@Map (key-type key-name, value-type value-name) -> (key-type, value-type) {
```

```
.
.
.
}
```

The first key-value pair defines the key and value types that form the input to the map function, and names these values so that they can be referred to in the body of the function. This is followed by an arrow and another key-value pair, defining the types of the output of the map function. Note that the output signature is *unnamed*, as these pairs are only produces at the end of the map function (**@All: mention something about emit() being similar to return() in that all code after it is unreachable?**).

The map function can include any number of calls to `emit()`, which outputs the resulting intermediate key-value pairs for use by the function defined in the `@reduce` section. The types of the values passed to the `emit()` function must agree with the signature of the output key-value pair as defined in the `@map` type signature. All output pairs from the map function are subsequently grouped by key by the framework, and passed as input to the `@reduce` function.

Currently, the only configuration available is for a file to be passed into the map function one line at a time, with the line of text being the value, and the corresponding line number as the key. This requires that the input key/value pair to the map function is of type `(int keyname,text valuename)`. Extending this to allow for other input formats is a future goal of the Hog language.

The following is an example of a complete `@Map` section for a program that counts the number of times each word appears in a set of files. The map function receives a single line of text, and for each word in the line (as delineated by whitespace), it emits the word as the key with a value of one. By emitting the word as the key, we can allow the framework to group by the word, thus calling the reduce function for every word.

2.4 @Reduce

The reduce function in a MapReduce program takes a list of values that share the same key, as emitted by the map function, and outputs a smaller set of values to be associated with another key. The input and output keys do not have to match, though they often do.

The setup for the reduce section is similar to the map section. However, the input value for any reduce function is always an iterator over the list of values associated with its key. The type of the key must be the same as the type of the key emitted by the map function. The iterator must be an iterator over the type of the values emitted by the map function.

```
@Reduce (key-type key-name, iter<value-type> value-name) -> (key-type, value-type)
{
```

```
.
.
.
}
```

As with the map function, the reduce function can emit as many key/value pairs as the user would like. Any key/value pair emitted by the reduce function is recorded in the output file.

Below is a sample at `@Reduce` section, which continues the word count example, and follows the @mapsample introduce in the previous section.

2.5 `@Main`

Fill this in! Hog.mapReduce()

Chapter 3

Lexical Conventions

3.1 Tokens

3.2 Comments

Block comments:

```
#{ these are block
   comments }#
```

Single-line comments are defined to be strings of text included between a '#' symbol on the left-hand side an a newline character ('\n') on the right-hand side.

3.3 Identifiers

3.4 Keywords

The reserved words of Hog are a superset of the reserved words of Java, since Hadoop scripts compile into runnable Java code. The following words are reserved for use as keywords, and may not be redefined by a programmer:

`abstract`	`case`	`default`	`file`
`assert`	`catch`	`do`	`final`
`bool`	`char`	`double`	`finally`
`boolean`	`class`	`else`	`float`
`break`	`const`	`enum`	`for`
`byte`	`continue`	`extends`	`goto`

`hadoop`	`long`	`reduce`	`text`
`if`	`map`	`return`	`this`
`implements`	`native`	`short`	`throw`
`import`	`new`	`static`	`throws`
`instanceof`	`package`	`strictfp`	`transient`
`int`	`private`	`super`	`try`
`interface`	`protected`	`switch`	`void`
`iter`	`public`		`volatile`
`list`	`real`	`synchronized`	`while`

@All: should we not say Java keywords are reserved, and instead rename any identifiers that could cause name conflicts? Feels weird to have a bunch of reserved words that are actually semantically meaningless for the language...

3.5 Constants

3.6 String Literals

3.7 Variable Scope

Hog implements what is generally referred to as lexical scoping or block scope. An identifier is valid within its enclosing block. The identifier is also value for any block nested within its enclosing block.

Chapter 4

Syntax Notation

Chapter 5

Types

5.1 Basic Types

The basic types of Hog include int (integer numbers, 64 bytes in size), real (floating point numbers, 64 bytes in size), bool (boolean values, true or false) and text (Strings, variable in size). Unlike most languages, Hog includes no basic character type. Instead, a programmer makes use of texts of size 1.

Implementation details Hogs primitive types are not so primitive. They are in fact wrappers around Hadoop classes. For instance, Hogs int type is a wrapper around Hadoop's IntWritableclass. The following lists for every primitive type in Hog the corresponding Hadoop class that the type is built on top of:

Hog Type	**Enclosed Hadoop Class**
int	IntWritable
real	DoubleWritable
bool	BooleanWrtiable
text	text???

5.2 Derived Types (Collections)

Derived types include list<T>, set<t>, multiset<t>, and iter<t>. The list<T> type is an ordered collection of objects of the same type. The set<T> is an unordered collection of unique objects of the same type. The multiset<T> is an unordered collection of objects of the same type, with duplicates allowed. The dict<K,V> is a collection of keyvalue pairs, where keys are all of the same type, and values are all of the same type (keys and values can be of different types from one another). The only types currently allowed within collections are primitive types, preventing such constructs as a list of lists. All collections allow for null entries.[1]

5.3 Conversions

[1] Note that for `set<T>`, only one `null` entry is allowed, and for `map<K,V>`, only one `null` key is allowed.

Chapter 6

Expressions

6.1 Operators

6.1.1 Arithmetic Operators

Hog implements all of the standard arithmetic operators. **Need to say something about using arithmetic operators with two operands of different type, and about using them with null values**.

Operator	Arity	Associativity	Precedence Level	Behavior
`+`	binary	left	0	addition
`-`	binary	left	0	minus
`*`	binary	left	1	multiplication
`/`	binary	left	1	division
`%`	binary	left	2 **??**	mod[1]
`++`	unary	left	3	increment
`--`	unary	left	3	decrement

6.1.2 Logical Operators

The following are the logical operators implemented in Hog. Note that these operators only work with two operands of type `bool`. Attempting to use a logical operator with an object of any other type results in an internal run-time exception (see §13.2). **@ALL: should this be a compile-time exception?**

Operator	Arity	Associativity	Precedence Level	Behavior
`or`	binary	left	0	logical or
`and`	binary	left	1	logical and
`not`	unary	right	2	negation

6.1.3 Comparators

The following are the comparators implemented in Hog (all are binary operations). **Need to say something about comparing two objects of different types, and null types**.

`<`	none	0	less than
`<=`	none	0	less than or equal to
`>`	none	0	greater than
`>=`	none	0	greater than or equal to
`==`	none	0	equal
`!=`	none	0	not equal

6.1.4 Assignment

There is one single assignment operator, '`=`'. Expressions involving the assignment operator have the following form:

```
lvalue = expression | PRIMITIVE | DERIVED
```

At compile-time, the compiler checks that both the result of the `expression` (or `PRIMITIVE` or `DERIVED`) and `lvalue` have the same type. If not, the compile throws a `TypeMistmatchException`. **@ALL: should this be a run-time exception?**

Chapter 7

Declarations

While it is not specified in the grammar of Hog, like many other programming languages, a user is only allowed to use variables/functions after they have been declared. **Another sentence of introduction**.

7.1 Type Specifiers

Every variable, be it a `primitive-type` or a `derived-type` has to be assigned a type upon declaration, for instance,

```
list<int> myList
```

Declares the variable `myList` to be a `list` of `int`s. And,

```
text myText
```

Declares the variable `myText` to be of type `text`.

7.2 Declarations

7.2.1 Null Declarations

If a variable is declared but not initialized, the variable becomes a ***null reference***, which means it points to nothing, holds no data, and will fail any comparison (see §6.1) for a discussion of how `null` affects comparisons and elementary arithmetic and boolean operations).

7.2.2 Primitive-type Variable Declarations

Variables of one of the primitive types, including `int`, `real`, `text` or `bool` are declared using the following patterns:

1. `variable-type variable-name` (a null declaration)

2. `variable-type variable-name = primitive-expr` (declaration with initialization)

When the first pattern is used, we say that the variable is ***uninitialized***, and has the value `null`. When the second pattern is used, we say that the variable is ***initialized***, and has the same value as the value of the result of the `primitive-assignment-expr`. The `primitive-assignmentexpr` must return a value of the right type, or the compiler will fail citing a syntax error. The `primitiveassignmentexpr` may contain an expression involving both other variables and unnamed raw primitives (e.g. 1 or 2), an expression involving only other variables or unnamed raw primitives, or a single variable, or a single unnamed raw primitive.

7.2.3 Derived-Type Variable Declarations

Derived-type variables are declared using the following patterns:

1. `variable-type variable-name`
2. `variable-type variable-name = derived-expr`
3. `variable-type variable-name(parameter-list)`
   `parameter-list -> parameter, parameter-list | parameter`

The first two patterns operate in essentially the same way as for primitivetype variables. When the first pattern is used, we say that the variable is ***uninitialized***, and has the value `null`. If a user attempts to use any type-specific operations (for instance, `size(myList)`) on an uninitialized variable, the program will through a runtime exception (see §13 for a discussion of exceptions). When the second pattern is used, the variable is ***initialized*** to the result of the `derivedexpr`.

Because derivedtype variables often have additional structure that needs to be defined at initialization, a third pattern is provided. In this pattern, the user can specify a list of ***parameters*** to initialize the object. For instance,

```
list<int> myList(5)
```

Specifies that `myList` should be initialized with five `null` values.

7.2.4 Function Declarations

See §2.2.

7.3 Initialization

Chapter 8

Statements

8.1 Expression Statement

8.2 Compound Statement

8.3 Flow-Of-Control Statements

@Paul: think about what you want here! Maybe include example for each? The following are the ***flow-of-control*** statements included in Hog:

```
if (expression) statement

if (expression) statement else statement

if (expression) statement elif (expression statement) ..  else
statement

switch(expression) statement
```

In the above statements, the ... signifies an unlimited number of elseif statements, since there is no limit on the number of elseif statements that can appear before the final else statement. In all forms of the if statement, the expression will be evaluated as a Boolean. If the expression is a number, then any nonzero number will be considered true and zero will be treated as false. In the second statement above, when the expression in the if statement evaluates to false, then the else statement will execute. In the third statement above with if, elseif and else statements, the statement will be executed that follows the first expression evaluating to true. If none of these expressions evaluate to true, then the else statement is executed.

The switch statement causes control to transfer to a statement depending on the matching case label. There can be an unlimited number of case labels within the switch statement, so that the switch will operate as such:

```
switch(expression) {
  case consant-expression : statment
  default : statement
}
```

An expression is passed in to the switch statement and then the flow of control will fall through the switch and the expression will then be compared to the constant expression next to each case label. When the switch expression matches the expression next to a specific case, the statement for that case is executed. If the flow of control falls to the bottom of the switch without finding an equality, then the default statement will be executed. The case constants are converted to the switch expression type. There cannot be two case expressions with the same value after conversion. In addition, there can only be one default label within each switch.

The above control statements can all be nested within each other.

8.4 Iteration Statements

Iteration statements signify looping and can appear in one of the two following forms:

```
while ( expression ) statement-list

for (expression ; expression ; expression ) statement-list

foreach expression in iterable-object statement-list
```

The following statement,

```
expression1
while(expression2) {
  statement
  expression3
}
```

is equivalent to,

```
for (expression1; expression2; expression3) statement
```

@Ben: clean up the above, needs to be a specific example, because it's not always true in this generic a fashion.

In the above while statement, expression1 will typically represent the initialization of a variable, then at each iteration through the while loop, the current Boolean value of expression2 is evaluated. Expression2 will be a test or condition. If expression2 evaluates to true, then iteration will continue through the while loop. Expression3 normally represents an increment step and its value changes at each step through the while. Expression3 could be a part of the condition tested in expression2. The first time expression2 evaluates to false, iteration through the loop ends and drops to the code that comes after the closing bracket of the while statement.

In the above for statement, expression1 is the initialization step, expression2 is the test or condition and expression3 is the increment step. At each step through the for loop, expression2 is evaluated. When expression2 evaluates to false, iteration through the loop ends.

When the foreach starts to execute, the iteration starts at the first element in the array or list and the statement executes during every iteration. The iteration ends when the statement has been executed for each item in the array (or list) and there are no items left to iterate through.

8.4.1 Example 1

```
int i = 0
while (i < 10) {
  print(i)
  i++
}
```

8.4.2 Example 2

```
for (int i = 0; i < 10; i++) {
  print(i)
}
```

8.4.3 Example 3

```
list<int> iList()
iList = [0,1,2,3,4,5,6,7,8,9]
foreach i in iList {
  print(i)
}
```

Chapter 9

Built-in Functions

Overview of built-in functions? How they are called on objects...

9.1 System-level Built-ins

Hog includes a number of systemlevel builtin functions that can be called from various sections of a Hog program. The functions are:

```
void emit(key, value)
```

This function can be called from the `@Map` and `@Reduce` sections in order to communicate the results of the map and reduce functions to the Hadoop platform. The types of the key/value pairs must match those defined as the output types in the header of each section.

```
void mapReduce()
```

This function can be called from the `@Main` section in order to initiate the mapreduce job, as definied in the `@Map` and `@Reduce` sections. Any Hog program that implements mapreduce will need to call this function in `@Main`.

```
void print(toPrint)
```

This function can be called from the `@Main` section in order to print to standard output. The argument must be a primitive type.

9.2 Object-level Built-ins

@ALL: introduction.

9.2.1 List

- `void sort()`

  Destructive?? function that sorts the items in the list in ascending order.

- `int len()`

  Returns an int with the number of elements in the list.

- `void append(itemToAppend)`

  Appends the object passed to the end of the list. The object must be of the same type as the list, or the operation will result in a **compile-time or run-time** exception.

- `<type> get(int index)`

  Returns the item from the list at the specified index.

9.2.2 text

The following function can be called on a `text` object:

```
list<text> tokenize(text fullText, text delimiter)
```

`tokenize()` can be called on a `text` object to tokenize it into a list of `text` objects based on the delimiter. The delimiter is not included in any of the `text` objects in the returned list.

Chapter 10

System Configuration

What facts about Hadoop instance does Hog need, and how to do you define them in a Hog program. Are there path variables that Hog depends on?

The user must set configuration variables in the `hog.rb` build script to allow the Hog compiler to link the Hog program with the necessary jar files to run the MapReduce job. The user must also specify the job name within the Hog source file.

HADOOP_HOME: absolute path of hadoop folder

HADOOP_VERSION: hadoop version number

JAVA_HOME: absolute path of java executable

JAVAC_HOME: absolute path of javac executable

HOST: where to job is rsynced to and run

LOCALMEM: how much memory for java to use when running in local mode

REDUCERS: the number of reduce tasks to run, set to zero for maponly jobs

Chapter 11

Parsing Tools

INCLUDE GRAPHIC

JFlex is a lexical analyzer generator for Java, and written in Java. JFlex is designed to work together with the LALR parser generator CUP.

Chapter 12

Linkage and I/O

INCLUDE GRAPHIC

12.1 Usage

To build and run a Hog source file there is an executable Ruby script `hog.rb` that automates the compilation and linking steps for the user.

Usage: `hog.rb [--hdfs|--local] job <job args>`

> `--hdfs`: if job ends in '.hog' or '.java' and the file exists, link it against the hadoop JARFILE and then run it on HOST.
>
> `--local`: run on local host.

12.2 Example

```
hog.rb local WordCountJob.hog input someInputFile.txt output ./someOutputFile.tsv
```

This runs the `wordCount` job in *local* mode (i.e. not on a Hadoop cluster).

Chapter 13

Exception Handling

Similar to other programming languages (Java, C++), Hog uses an exception model in which an exception is thrown and can be caught by a catch block. Code should be surrounded by a try block and then any exceptions occurring within the try block will subsequently be caught by the catch block. Each try block should be associated with at least one catch block. However, there can be multiple catch blocks to handle specific types of exceptions. In addition, an optional finally block can be added. The finally block will execute in all circumstances, whether or not an exception is thrown. The structure of exception handling should be similar to this, although there can be multiple catch blocks and the finally block is optional:

```
try {
  statement-list
} catch (exception) {
  statement-list
} finally {
  statement-list
}
```

Because the proper behavior of a Hog program is dependent on resources outside of the language (i.e. the proper behavior of the users Hadoop software), there are more sources exceptions in Hog than most general purpose languages. These sources can be divided into three categories: ***compiletime exceptions***, ***internal runtime exceptions*** and ***external runtime exceptions***.

13.1 Compile-time Exceptions

The primary cause of most compiletime exceptions in Hog are syntax errors. Such errors are unrecoverable because it is impossible for the compiler to properly interpret the user program. Some compilers for other languages offer a limited amount of compiletime error correction. Because Hog programs are often designed to process gigabytes or terabytes of data at a time, the standard Hog compiler offers no compiletime error correction. The assumption is that a user would rather retool their program than risk the chance of discovering, only after hours of processing, that the compilers has incorrectly assumed what the user meant. The following are Hog compiletime exceptions:

Add descriptions for each exception

ProgramStructureException

NoSuchVariableException

NoSuchMethodException

UnsupportedOperationException

RedundantDeclarationException

13.2 Internal Run-time Exceptions

Internal runtime exceptions include such problems as I/O exceptions (i.e. a specified file is not found on either the users local file system or the associated Hadoop file system), type mismatch exceptions (i.e. a program attempts to place two elements of different types into the same list) and parsing exceptions. **Another sentence or two**. The following are Hog internal rumtime exceptions:

Add descriptions for each exception

FileNotFoundException

FileLoadException

ArrayOutOfBoundsException

IncorrectArgumentException

TypeMismatchException

HogMapFunctionException

HogReduceFunctionException

NullPointerException

NumberFormatException

ArithmeticException

IOException

InterruptedException

ParseException

13.3 External Run-time Exceptions

Hog programs are primarily meant to be run on a Hadoop cluster, and therefore a third source of exceptions for the language comes from errors associated with the cluster. If the cluster becomes corrupted in the middle of processing a Hog script (**can we say anything but "the hog script dies?" @Kurrydoes Hadoop have a single pointoffailure? If so, we should say something about it here**). The following are external runtime exceptions (**@kurry, any input on these?**):

Add descriptions for each exception

HadoopCompilationException

HadoopFailuerException

HadoopExecutionException

HadoopMapFunctionException

HadoopReduceFunctionException

Chapter 14

Grammar

```
if ( expression ) statement
if ( expression ) statement else statement
if ( expression ) statement elseif ( expression ) statement... else statement

switch ( expression ) statement if ( expression ) statement
if ( expression ) statement else statement
if ( expression ) statement elseif ( expression ) statement... else statement

switch ( expression ) statement

while ( expression ) statement
for ( expression; expression; expression ) statement
foreach ( expression in array or list) statement
@Functions
returntype functionname1 (parameterlist) {
exprlist }
returntype functionname2 (parameterlist) { exprlist
}
@Map: (type keyname, type valuename) > (type, type) {
.
.
.
}
```