Pipeline Architectures

Software Architecture

March 7, 2022

Brae Webb

1 Brief

This week we are looking into architectures which follow the pipeline architectural style. Specifically, this week we will:

1. Briefly recap the concepts of pipeline architectures.
2. Explore the advantages of using bash as a well-implemented pipeline architecture using the Knuth/McIlroy example.
3. Explore whether compilers are a suitable pipeline architecture.
4. Briefly look at one of the primary modern applications of the pipeline style, MapReduce.

2 Case Study: Compilers

An interesting case study of the pipeline architecture is a compiler.[1] As a foundational technology, compilers have undergone rigorous refinement and are perhaps the most well studied type of software. Modern compilers have well-defined modular phases as illustrated by Figure 1, each phase of a compiler transforms the representation of the program until the target program is produced.

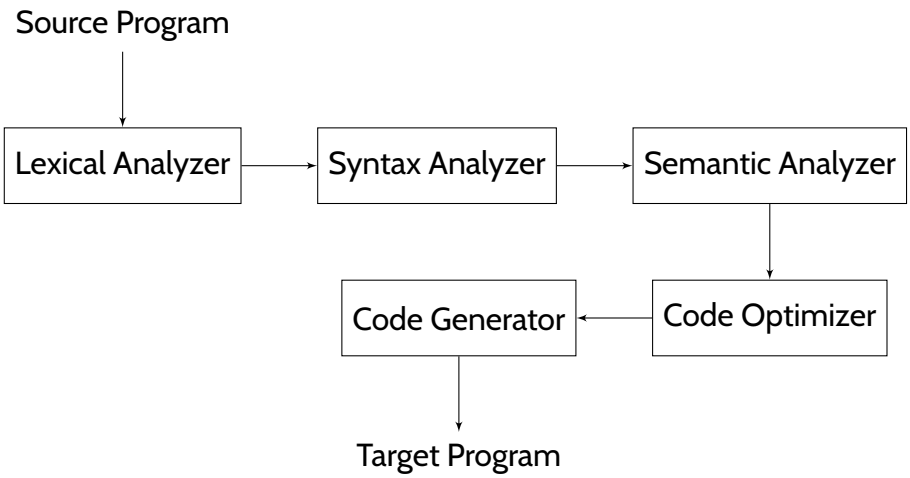

Figure 1: Typical phases of a compiler.

However, a compiler is not well suited to use a pipeline architecture. In general, the modules of a pipeline architecture should be independent of their input source. This is not the case in compilers, as each phase relies on the completion of the previous phase. As a result, the input dependencies of a compiler make it too restrictive for a true pipeline architecture.

Instead, compilers are often built as a hybrid of a pipeline architecture and the *Blackboard Architecture*. The blackboard architecture consists of;

[1] You don't need to understand the phases of a compiler — two data structures, the Symbol Table and AST, are transformed in each compiler phase.

- a knowledge base, the 'blackboard',
- knowledge sources which use and update the knowledge base, and
- a control component to coordinate the operation of knowledge sources.

In modern compilers, the data which would be passed through pipes, the Symbol Table and AST, are used and updated by each phase. They are subsequently used as the 'blackboard'. Each phase is considered a knowledge source which uses the knowledge base to transform and update the knowledge base. Finally, in this hybrid, the control component is not required as the sequence of phase execution in a pipeline coordinates operation. Figure 2 illustrates this proposed architecture. Of course, there are many compilers out there, many of them deviate from this architectural hybrid.

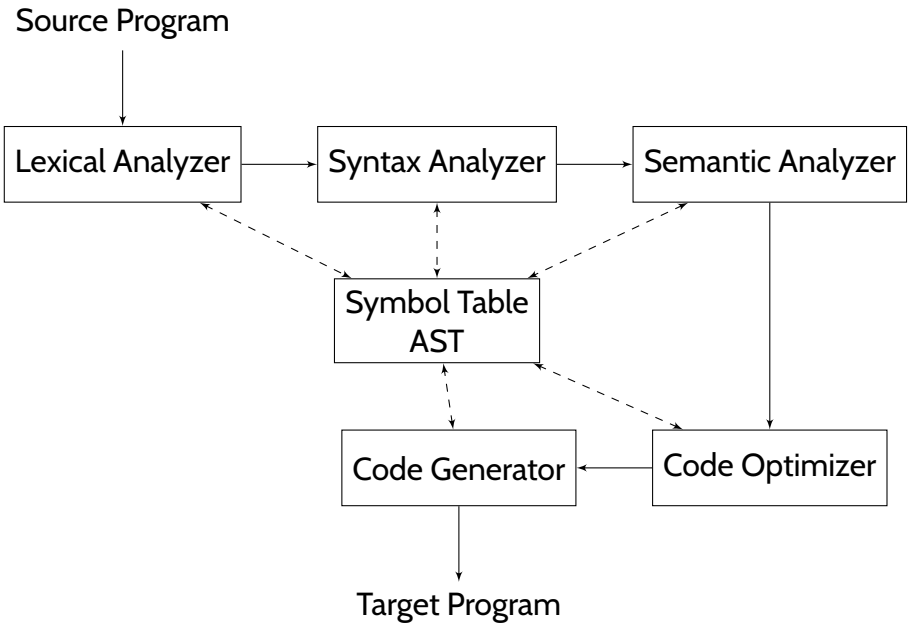

Figure 2: Modern phases of a compiler.

3 Case Study: MapReduce

One of the more prevalent uses of the pipeline architecture is the MapReduce pattern. The MapReduce pattern was discovered in 2004 as a solution to the challenges which Google faced managing their search index [1].[2] MapReduce affords impressive parallelism inherit to the programming pattern.

The two key ideas of MapReduce, are *map* and *reduce*. Below are the generic types of the *map* and *reduce* functions in functional programming.

```
1 map : (τ1 → τ2) → τ1Seq → τ2Seq
2 map f xs
3 reduce : (τ2 → τ1 → τ2) → τ1Seq → τ2 → τ2
4 reduce f xs initial
```

If you're unfamiliar with this notation, the rough English translation is:

map The parameters of the *map* function are:

[2] Although the pattern was in use prior to their work[2]

(a) A function, f, which takes a parameter of type τ_1 and returns a type τ_2.

(b) A sequence of elements of type τ_1.

The return type of the *map* function is a sequence of elements of type τ_2.

reduce The parameters of the *reduce* function are:

(a) A function, f, which takes two parameters, the first of type τ_2 is the current accumulator value, the second of type τ_1 is the current value in the input sequence, and returns a τ_2, the new accumulator value.

(b) A sequence of elements of type τ_1.

(c) An initial accumulator value of type τ_2

The return type of the *reduce* function is the type τ_2.

The code snippet below uses the map and $reduce$ functions to perform the operations of the above bash example. One important thing to note about the example below is the map operation on line 11. Each application of the lambda function within the map operation is completely independent and could, in theory, be executed simultaneously.

```
1  contents = read("assignment.py")
2
3  # filter relevant lines by rebuilding the list
4  contents = reduce(λ xs x →
5                    if x.contains("hack")
6                      then x + xs
7                      else xs,
8                  contents,
9                  [])
10
11 # use map to count occurrences of word
12 contents = map(λ line →line.count("hack"), contents)
13
14 # use reduce to sum list of counts
15 contents = reduce(λ total curr →total + curr, contents, 0)
16
17 write("code-quality.txt", contents)
```

So by design, code written in this pattern can process data simultaneously. Tools such as Hadoop[3] are able to take advantage of this to distribute computation automatically.

Using the terminology of a pipeline architecture, what filters do the *map* and *reduce* operators roughly correspond to?

How would you improve the efficiency of the code snippet above?

[3] https://hadoop.apache.org/

References

[1] J. Dean and S. Ghemawat, "Mapreduce: Simplified data processing on large clusters," in *OSDI'04: Sixth Symposium on Operating System Design and Implementation*, (San Francisco, CA), pp. 137–150, 2004.

[2] D. J. DeWitt and M. Stonebraker, "Mapreduce: A major step backwards." https://dsf.berkeley.edu/cs286/papers/backwards-vertica2008.pdf, January 2008.