Alexander Smith

[Company name] | [Company address]

[Document title]

[Document subtitle]

# Abstract

As the field of computer science expands to include ever more methodologies and specialities, from mathematical proofs to artificial intelligence, weather modelling to financial modelling, general purpose programming languages end up being the Swiss army knife of tools, ok at everything, but poor at any one specific job. One such speciality is data processing, covering everything from initial extract, transform and load workloads, to big data exploration and back to tabulation of complex results.

Further, general purpose languages offer many solutions to data processing, often encouraging the user to read an entire file into memory and processing it many times into new memory structures. This methodology if often good enough for “simple” tasks, but rarely scales well. On the other hand, efficient streaming or more modern generator paradigms can be harder to implement when the programmer is not familiar with them.

“Hoshie Lang” is a new programming language that specialises in processing data. It has the following goals:

1. Enable the programmer to succinctly define what they want to happen in their data process.
2. Provide the programmer with a built-in library of common algorithms and functions used to process and analyse data.
3. Provide a pleasant programming experience in an IDE.
4. Encapsulate the execution code generation in a way that can be extended and adapted in the future to ensure Hoshie Lang can be executed on a variety of platforms.
5. Remove the need for the programmer to be familiar with memory efficiency and streaming / generator techniques.

## Succinct Programming

The Hoshie Lang programming language is based around a “Pipeline” analogy. The programmer is encouraged to think of the data as “flowing” through a pipe (one row at a time), which in turn is made up of:

* **Activities**: A section in the pipe that alters the data in some way, typically:
  + Altering the content of a row.
  + Augmenting or Truncating the format of a row.
  + Removing specific rows.
  + Appending new rows.
* **Sensors**: A section in the pipe that observes the data as it passes that point in the pipeline, these are typically **reducers** which calculate scalar values based on the data it has seen up to that point in time.

The language is **declarative**, which places a natural emphasis on breaking down the entire process into a smaller set of independent **actions**, which in turn can then be used in specific data processing pipelines.

Similarly, **sensors** can be defined and placed along the pipeline to observe the state of process at any given time.

Rectangle = Actions  
Ellipses = Sensors

## Built-in Activities and Sensors

The language includes a set of predefined functions that are used to define the actions and sensors. For example, some common actions are:

* **Filter**: Allow a subset of rows to flow through the pipe.
* **Map**: Maps a specific row structure to a new row structure.
* **FirstN**: Only allow N rows to pass through.

And some common Sensors would be:

* **Count**: A simple row count.
* **Max**: The max value of a nominated field.
* **Mean**: The mean value of a nominated field.

## IDE

For languages to become popular and be usable, they need to have good integrated development environment support as this lowers the barrier of entry when learning the language and can increase the programmer’s productivity by an order of magnitude.

## Encapsulated Code Generation

One of the visions for Hoshie Lang is to be able to process data on a variety of platforms. For example:

* Small data could be processed within a web browser.
* Medium data could be processed within a middleware application, NodeJS, Python or Java.
* Large data could be processed in a distributed computing cluster like Hadoop or Spark.

By ensuring the code generation is well encapsulated adding new code generators can be a relatively easy exercise, preferably by domain experts in those target platforms. For this paper, the reference implementation is targeting JavaScript which can cover the small to medium data sizes above.

## Efficient Runtime

Having an efficient runtime will be an important part of Hoshie Lang. In the reference JavaScript implementation, a streaming based runtime is used, making heavy use of JavaScript generators. This effectively means that data rows are “pulled” from the end of the pipe, triggering the upstream actions to do “just enough” work for each row to be created.

For example, if the source dataset is 1 billion rows in length, but the processing pipeline contains a “FirstN” action (e.g., one which only lets the first 100 rows through), then the bulk of source dataset will never be processed.

## Conclusion

There are a lot of moving parts when creating a new programming language, all of which play a key part to the overall usability of the language.

There are also a lot of tools, libraries, and IDEs available to assist in the task. This paper will take an in depth look at these and how they can be utilised to create a final product. Specifically:

* Language design.
* Compiler compilers, lexers and parsers.
* IDE integration and APIs
* Data processing and streaming techniques
* Documentation, unit test and build tools

# Declaration

# Acknowledgements

Table of Contents

[Abstract 1](#_Toc67743466)

[Succinct Programming 2](#_Toc67743467)

[Built-in Activities and Sensors 2](#_Toc67743468)

[IDE 2](#_Toc67743469)

[Encapsulated Code Generation 3](#_Toc67743470)

[Efficient Runtime 3](#_Toc67743471)

[Conclusion 3](#_Toc67743472)

[Declaration 4](#_Toc67743473)

[Acknowledgements 4](#_Toc67743474)

[Introduction 7](#_Toc67743475)

[Analysis 10](#_Toc67743476)

[Syntax Design 10](#_Toc67743477)

[Compiler 10](#_Toc67743478)

[Integrated Development Environment 11](#_Toc67743479)

[Runtime Environment 12](#_Toc67743480)

[Design 13](#_Toc67743481)

[Syntax 13](#_Toc67743482)

[Compiler 13](#_Toc67743483)

[Design Conclusion 14](#_Toc67743484)

[Implementation 15](#_Toc67743485)

[Development Environment 15](#_Toc67743486)

[Hoshi Lang Parsing 16](#_Toc67743487)

[Hoshie Lang Grammar 17](#_Toc67743488)

[Creating the Abstract Syntax Tree 21](#_Toc67743489)

[Navigating the Abstract Syntax Tree 21](#_Toc67743490)

[Integrated Development Environment 21](#_Toc67743491)

[Contribution Points 21](#_Toc67743492)

[Evaluation 22](#_Toc67743493)

[Conclusions 22](#_Toc67743494)

[Appendix 22](#_Toc67743495)

[References 22](#_Toc67743496)

[References 22](#_Toc67743497)

# Introduction

The overall goal of “Hoshie Lang” was to create a new programming language that focused on the task of Data Processing.

Having spent time working on several class assignments, intern projects and hackathons there was a common need to load one or more raw data files and perform some initial processing and analytics to help understand, interrogate and finally format the data before we could attend to the actual job or assignment at hand. Invariably this step was implemented in a very ad hoc way as it was a secondary concern, which ended up with the developer rewriting similar code in different circumstances and often with different programming languages.

Further, while programming languages usually provide plenty of support to get the job done, the programmer will often take the easiest approach which may be fine for smaller datasets but will fail to scale with the data efficiently. For example, they may load the entire dataset into memory and use a FOR loop to iterate over it, when a streamed reader / writer would be more efficient but more complex to code.

Finally, with any quick ad hoc code comes the additional risk of introducing subtle bugs, especially when the task appears to be secondary to the programmer.

Hoshie Lang attempts to solve these issues by presenting a data processing centric programming language. By design it encourages the programmer to define **what** data processing and analytics to perform on the data, without having to specify the **how** to do it. For example, Hoshie Lang has no keywords to “LOOP” or “STREAM” data, instead it has an abstraction of a “Pipeline”, one where data flows from the top of the pipe to the bottom, while passing through a series of “Activities” and “Sensors”, where an “Activity” alters the data in some way and a “Sensor” observes the data as it passes.

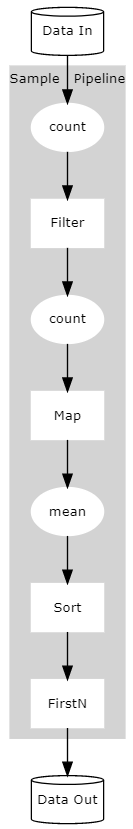


Figure 1

An example of a Simple Data Pipeline

This conceptually shows how the data flows from top to bottom, through a data processing “Pipeline”.

Each “Activity” is represented by a rectangle.

Each “Sensor” is represented by an ellipse

**Activity**: Typically changes the format of data (MAP) or the content of the dataset (FILTER).

**Sensor**: Observes the data as it passes that point in the “Pipeline”, typically reducing the data to a single value or set of values (COUNT, MEAN).

To achieve this goal the project was broken down into four distinct steps:

* Designing and specifying the syntax of the new language.
* Creating a lexer / parser for this new syntax.
* Ensuring Integrated Development Environment (IDE) support exists for the new language.
* Create an encapsulated Code Generation layer that initially targets NodeJS / Browsers (JavaScript), but could be extended in the future to target other platforms (Python, Spark etc.).

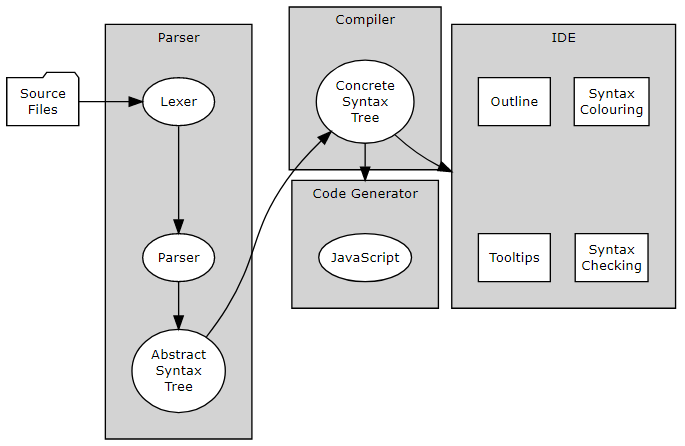


Figure 2 - Overview

# Analysis

The overall project can be broken down to several key functional areas:

* Syntax Design: The design of the language - should be succinct, but easy to read / write.
* Compiler: The compiler will convert the new language into a format that can be executed.
* Integrated Development Environment: Having good IDE integration is a key factor to any language’s usability.
* Runtime environment: Where can the compiled program be executed?

## Syntax Design

* Functional or OOP?
  + Pros:
  + Cons:
* Immutable or Mutable data?
  + Pros:
  + Cons:
* Declarative or Procedural?
  + Pros:
  + Cons:
* Scoping rules:
  + Files
  + Functions
* Visibility rules:
  + Forward declarations
* Operators
* Declarations
* Actions
* Built-in functions

## Compiler

There are some common tools to assist in compiler development, these are known as “compiler compilers”:

* Antlr?
  + Pros:
  + Cons:
* Flex / Bison?
  + Pros:
  + Cons:
* Lex / Yacc?
  + Pros:
  + Cons:

## Integrated Development Environment

* Bespoke: A specific IDE developed for the language in question.
  + Pros:
    - Can cater to the language very well
    - Can have deep integration with the language
  + Cons:
    - A lot of work
    - Developers do not like learning new IDEs
* Visual Studio: Microsoft’s IDE, primarily focused on Windows development
  + Pros:
  + Cons:
* XCode: Apples IDE, primarily focused on OSX + IOS development
  + Pros:
  + Cons:
* Android Studio: Googles IDE, primarily focused on Android and Chrome OS development
  + Pros:
  + Cons:
* IntelliJ: IDE focusing on Java Development
  + Pros:
  + Cons:
* Eclipse: Designed to be an extensible IDE written in Java
  + Pros:
  + Cons:
* Atom: “A hackable text editor for the 21st Century”
  + Pros:
  + Cons:
* VS Code: Designed from the start to be an unopinionated modern and extensible editor / IDE written in TypeScript. Borrowed heavily from Atom’s ideology
  + Pros:
  + Cons:

## Runtime Environment

* Native Machine Code:
* Java Virtual Machine:
* JavaScript Environment:
  + Browser
  + NodeJS
  + Deno

# Design

## Syntax

The key areas for the language are:

* Types:
  + Primitive: boolean, number, string
  + Complex: A named “set” of Primitives / Complex items
  + Arrays: A “list” of Complex or Primitive items (limited to one type)
* Expressions: A single entity that can be evaluated to a value:
  + Boolean: true, false, equality, greater, less, and, or, not etc.
  + Numeric: Plus, minus, multiplication, division, etc.
  + String: Concatenation
* Functions: An encapsulated set of expressions:
  + Gated inputs (parameters)
  + Internals should be private.
  + Single return value.
  + Should be pure.
* Declaration: A named Type, Expression or Function
* Scoping:
  + Files: Declarations inside a file will only be visible within that file unless a declaration is specifically exported.
  + Functions:
    - Declarations within the function are only visible within that function.
    - Declaration outside the function will not be visible within that function.
    - Parameters will be used to pass in expressions / declarations.
    - Return value will be used to expose a result from the function.
    - Functions are expected to be pure.
* Built-in Functions: The language will have a set of data processing functions built into the language. It will also have built-in support for unit testing.

## Compiler

* Lexer: The set of known tokens, keywords, and operators
* Grammar: Defines how the tokens can be structured in a high-level way

## Design Conclusion

Primary development language would be TypeScript and the targeted runtime environment would NodeJS / Browser (JavaScript). The key factors in this decision were:

* Antlr supports generating JavaScript (and experimental TypeScript) parsing and visitor code.
* VS Code extension authoring is written in TypeScript.
* NodeJS and Browser environments are cross-platform and widely available.

# Implementation

This section describes the development environment, tools and programming languages used to develop Hoshie Lang.

## Development Environment

The following is a description of the development environment used, given the nature of the project setting up an optimised environment served as valuable research as well as expediting actual development.

The following development tools were used:

* **NodeJS**: Cross-platform JavaScript runtime environment for executing JavaScript outside of a web browser. Has a large ecosystem of third-party libraries and a popular package manager to define versioned dependencies - <https://nodejs.org/en/>
* **TypeScript**: An extension to the JavaScript language, TypeScript adds static typing and transpiling capabilities making large scale development of JavaScript more robust - <https://www.typescriptlang.org/>
* **Antlr**: A “compiler compiler” which takes Lexer, and grammar definition files and generates parsers, abstract syntax trees with associated walkers and visitors - <https://www.antlr.org/>
* **Mocha**: JavaScript test framework - <https://mochajs.org/>
* **Chai**: A JavaScript assertion library, commonly used in conjunction with Mocha - <https://www.chaijs.com/>
* **faker**: A JavaScript library for generating fake test data - <https://github.com/Marak/Faker.js>
* **rimraf**: Emulated the "rm -rf" command line in JavaScript. Used for running "clean" commands - <https://github.com/isaacs/rimraf>
* **npm-run-all**: Allows sequential and parallel execution of "npm run" commands. Used for "build-all" type commands - <https://github.com/mysticatea/npm-run-all>
* **watch**: Monitors a folder for changes, used to detect when files change so they can be automatically recompiled - <https://github.com/mikeal/watch>
* **pandoc**: A universal document converter, used to convert markdown to Microsoft Word formats (and back again), allow initial report writing to be done in markdown - <https://pandoc.org/>
* **Visual Studio Code**: Development environment - <https://code.visualstudio.com/>, including the following extensions:
  + **ANTLR4 grammar syntax support**: IDE support for Antlr lexer and grammar files - [https://marketplace.visualstudio.com/items?itemName=mike-lischke.Visual Studio Code-antlr4](https://marketplace.visualstudio.com/items?itemName=mike-lischke.vscode-antlr4)
  + **ESLint**: IDE support for TypeScript linting tools - [https://marketplace.visualstudio.com/items?itemName=dbaeumer.Visual Studio Code-eslint](https://marketplace.visualstudio.com/items?itemName=dbaeumer.vscode-eslint)

## Hoshi Lang Parsing

When developing a new programming language, the first hurdle is to be able to read and understand the files that contain the “Hoshie Lang program” as written by the programmer.

This first step will take the structured text documents and parse them into a logical data representation that can then be used in the next step of the overall process. A side effect of this parsing is that any text that cannot be parsed would be considered a “syntax error”. It is worth noting that it is possible to have syntactically correct code which cannot execute due to other errors (referencing a declaration which has not been declared or is the wrong type for example). Syntax errors are the most basic error the programmer can make and can prevent the program from even being parsed completely. It is common for these types of errors to be caught quickly in an IDE (as the user types) and highlighted so the programmer can fix them quickly.

The Hoshie Lang parser was developed using a tool called ANTLR - *“(****AN****other* ***T****ool for* ***L****anguage* ***R****ecognition) is a powerful parser generator for reading, processing, executing, or translating structured text or binary files. It's widely used to build languages, tools, and frameworks. From a grammar, ANTLR generates a parser that can build and walk parse trees.”* (ANTLR Homepage n.d.).

At a high level the ANTLR tool helps abstract the Hoshie Lang language into a formal definition, commonly known as the “Grammar”, from which it will generate the actual code that converts the text into a usable data structure. This data structure is referred to as an “Abstract Syntax Tree” or AST, as the name suggests it is tree shape structure where each node represents part of the program. Further, ANTLR will also create helper functions for navigating and interrogating the AST, in our case these helpers are referred to as “Visitors” and provide a convenient way to traverse and perform actions on each node in the AST.

### Hoshie Lang Grammar

The formal definition of the Hoshie Lang Grammar is written in the ANTLR “meta-language” and are broken into two distinct parts:

* Lexer
* Parser

This is akin to how a human might read a book, humans don’t read a book one character at a time, but rather they recognise the individual words and process them as a stream that forms a sentence. The meaning of the sentence is then derived from the order and relationship of the words.

The process of grouping the individual characters and symbols into “Tokens” is known as Lexical Analyses (or tokenizing), in ANTLR this is referred to as the “Lexer”. The syntax structure for ANTLR lexers can be found at:  
<https://github.com/antlr/antlr4/blob/master/doc/lexer-rules.md> (Lexer Rules n.d.).

The second part of the process is the “Parser”, the parser takes the stream of tokens and organises it into a tree structure, commonly known as a “parse tree” or “abstract syntax tree” (AST). The syntax structure for ANTLR parsers can be found at: <https://github.com/antlr/antlr4/blob/master/doc/parser-rules.md> (Parser Rules n.d.)

The AST data structure for a syntactically correct program will precisely represent the entire program and can be used for many different tasks without having to reparse the original program. For example the AST could be used to:

* Colourise the program in an IDE.
* Reformat the original text in the program to a consistent structure, based on a set of rules (tabs or spaces, “=” versus “ = “ etc.).
* Generate executable code (ASM, c++, JavaScript etc.).
* Generate HTML documentation.

#### Hoshie Lang Lexer Definition

The source code the Hoshie Lang Lexer can be located in the sources at “grammar/HLLexer.g4”. It consists of the following main sections (all examples are a small section of the actual code):

* Symbols: A named list of valid symbols, for example:

Comma: ',';

Assign: '=';

Not: '!';

Multiply: '\*';

* Keywords: A named list of keywords, for example:

Boolean: 'boolean';

Number: 'number';

String: 'string';

Return: 'return';

* Functions: Built-in function names:

Length: 'length';

Generate: 'generate';

Random: 'random';

UTest: 'utest';

* Literals: Covers literal values, like boolean, numbers, strings etc.:

NullLiteral: 'null';

BooleanLiteral: 'true' | 'false';

DecimalLiteral

  : DecimalIntegerLiteral '.' DecimalDigit\* ExponentPart?

  | '.' DecimalDigit+ ExponentPart?

  | DecimalIntegerLiteral ExponentPart?

  ;

fragment DecimalDigit: [0-9];

fragment DecimalIntegerLiteral: '0' | [1-9] DecimalDigit\*;

* Identifiers: Hoshie Lang identifiers (must start with a character or “\_” and may consist of characters, numbers and “\_”:

Identifier: IdentifierStart IdentifierPart\*;

fragment IdentifierPart

  : [\_]

  | [a-zA-Z]

  | [0-9]

  ;

fragment IdentifierStart: [\_] | [a-z];

* Strings: Single, double quote strings and escape character definitions
* Other: Whitespace and comment markers

#### Hoshie Lang Parser Definition

The source code the Hoshie Lang Parser can be located in the sources at “grammar/HLParser.g4”. The entire grammar is complex and heavily interlinked, but some of the highlights are:

* Entry point – the root node of the AST:

program: fileElements? EOF;

* File Statement – the main parts of each file:

fileStatement

  : block

  | typeStatement

  | variableStatement

  | actionStatement

  | importStatement

  | exportStatement

  | emptyStatement

  ;

* Assignments – typically assigning an identifier to an expression or function:

variableStatement: variableDeclaration eos;

variableDeclaration: Identifier variableInitialiser;

variableInitialiser

  : '=' singleExpression (As singleTypeExpression)?

  ;

* Single Expressions – these appear in many places, for example in function calls, on the right hand side of expressions and in equality statements:

singleExpression

  : singleExpression arguments

  | '!' singleExpression

  | singleExpression ('\*' | '/' | '%') singleExpression

  | singleExpression ('+' | '-') singleExpression

  | singleExpression ('<' | '>' | '<=' | '>=') singleExpression

  | singleExpression ('==' | '!=') singleExpression

  | singleExpression ('&&' | '||') singleExpression

  | identifier

  | literal

  | arrayLiteral

  | arrowFunction

  | keyword arguments

  ;

#### ANTLR Command Line Tool

The ANTL command line tool is used to convert the lexer and grammar files into actual source code which can be used within our application. Specifically, it is used to generate the following key pieces:

* Lexer / Tokenizer
* Parser
* Error Listener
* Visitor

The command line used for Hoshie Lang:

antlr4  
 -Dlanguage=JavaScript   
 -o ./src/hlcc   
 -lib ./src/hlcc/grammar   
 -visitor   
 ./grammar/HLLexer.g4 ./grammar/HLParser.g4

Where:

* -Dlanguage: Specifies which programming language should be targeted for the output, in this case JavaScript.
* -o: Specifies output directory where all output is generated.
* -lib: Specifies location of grammars, tokens files.
* -visitor: Enable generation of parse tree visitor
* <input files>

### Creating the Abstract Syntax Tree

Having generated the four key pieces, they can now be integrated into our application and be used to parse Hoshie Lang files. The pseudo code to convert the Hoshie Lang text file into an Abstract Syntax Tree is:

charStream = InputStream(“Hoshie Lang Program”)

lexer = HLLexer(charStream)

tokens = CommonTokenStream(lexer)

parser = HLParser(tokens)

ast = parser.program()

1. Create a character stream from the hoshie lang text or file.
2. Create a new Hoshi Lang lexer instance from the character stream.
3. Convert the lexer to a stream of tokens.
4. Create a new instance of the parser from the token stream.
5. Get a reference to the root node in the abstract syntax tree.

The actual code can be viewed in: src/hlcc/parser.ts

### Navigating the Abstract Syntax Tree

## Integrated Development Environment

A critical part of any programming language is good support for associated development tools, VS Code is currently one of the most popular cross platform, general purpose programming IDEs available. https://code.visualstudio.com/

### Contribution Points

VS Code has many ways it can be extended to support new features and languages. These capabilities are defined using “Contribution Points” within the extensions **package.json** file, in a specific section called **contributes**. https://code.visualstudio.com/api/references/contribution-points

* **languages**: Every extension will need to declare some global information about itself, this includes:
  + **id**: A unique ID associated with the extension (hoshie)
  + **extensions**: File extension IDs that VS Code can use to associate a file with the extension (.ho)
  + **configuration**: Some general rules for the language to help the editor handle common operations (./configuration.json) and includes - [link](https://code.visualstudio.com/api/language-extensions/language-configuration-guide):
    - Comment toggling
    - Bracket’s definition
    - Auto closing
    - Auto surrounding
    - Folding
    - Word pattern
    - Indentation Rules
* **grammars**: VS Code uses [TextMate grammar definitions](https://macromates.com/manual/en/language_grammars) for its default syntax colouring. This contribution point associates hoshie lang with a Text Mate language file **./syntaxes/hoshie.tmLanguage.json**.
* **commands**: scasc

# Evaluation

# Conclusions

# Appendix

# References

* Grammar naming conventions and layout based on:
  + https://github.com/antlr/grammars-v4/blob/master/javascript/ecmascript/JavaScript/ECMAScript.g4
  + https://github.com/antlr/grammars-v4/blob/master/javascript/javascript/JavaScriptParser.g4
  + https://github.com/antlr/grammars-v4/blob/master/javascript/javascript/JavaScriptLexer.g4
* Visual Studio Code Syntax Language file bases on
  + https://github.com/microsoft/Visual Studio Code/blob/master/extensions/javascript/syntaxes/JavaScript.tmLanguage.json

# References

n.d. *ANTLR Homepage.* https://www.antlr.org/.

n.d. *Lexer Rules.* https://github.com/antlr/antlr4/blob/master/doc/lexer-rules.md.

n.d. *Parser Rules.* https://github.com/antlr/antlr4/blob/master/doc/parser-rules.md.