

Практика 2: Пример архитектуры — bash

22.02.2018

1. Enterprise Fizz-Buzz

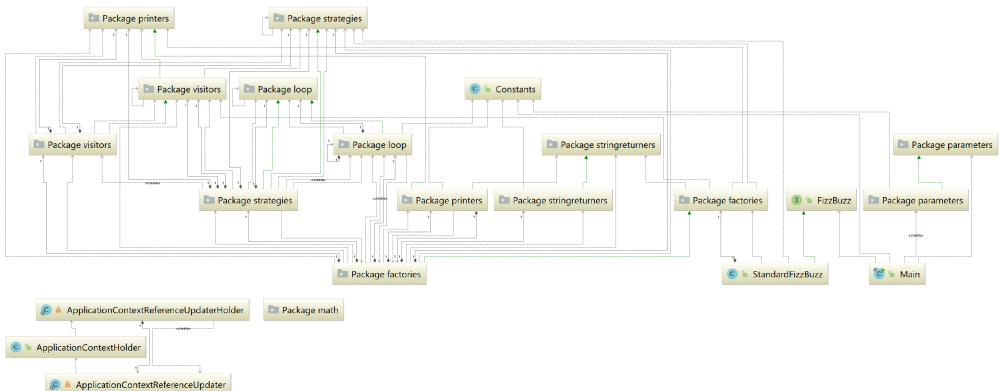
Сегодняшняя пара будет про примеры архитектур реальных приложений. Начнём мы с самого реального приложения — FizzBuzz Enterprise Edition, созданного серьёзными бизнесменами для серьёзных бизнес-задач. Собственно, серьёзная бизнес-задача такая:

Для чисел от 1 до 100:

- если число делится на 3, вывести “Fizz”;
- если число делится на 5, вывести “Buzz”;
- если число делится и на 3, и на 5, вывести “FizzBuzz”;
- во всех остальных случаях вывести само число.

Правильное архитектурное решение можно найти тут: <https://github.com/EnterpriseQualityCoding/FizzBuzzEnterpriseEdition>.

Вот обзорная диаграмма, описывающая в общих чертах структуру этого проекта, которая была автоматически построена IntelliJ IDEA:



По данным OpenHub.net, реализация содержит 1663 содержательные строки кода и всего 40 строк комментариев (ужасно). OpenHub.net оценивает стоимость разработки этого проекта в 18K долларов.

Проект использует Spring Framework, активно используются паттерны “Фабрика”, “Стратегия”, “Посетитель”. Достигнуто аккуратное разделение ответственностей, например, за исполнение цикла от 0 до 100 отвечают классы LoopContext и LoopRunner. Границы цикла, к несчастью, вынесены в глобальные для всего приложения константы, а не читаются из конфигурационного файла базы данных. Зато используется важный для объектно-ориентированной архитектуры принцип инверсии зависимостей — классы не знают ничего про другие конкретные классы, а содержат ссылки только на интерфейсы. Соответственно, практически каждый класс имеет свой интерфейс, через который с ним могут взаимодействовать остальные, и свою фабрику, которая может создать объект и вернуть его как ссылку на интерфейс, не раскрывая клиентам наличие даже конструктора у класса. Поэтому на диаграмме видны пакеты с одинаковыми названиями — один для классов-реализаций, другой с интерфейсами.

2. Архитектура bash

Теперь перейдём к более “настоящему” проекту — командному интерпретатору Bash. Это не то чтобы хороший пример хорошей объектно-ориентированной архитектуры (он был написан на C, причём ещё в те времена, когда объектно-ориентированная архитектура была не очень модна), к тому же его архитектура не очень подробно документирована. Но нам этот проект интересен, во-первых, в свете домашки, а во-вторых тем, что его часто (ну, относительно) используют в исследованиях по архитектуре как “подопытного кролика”. Суммарно bash имеет около 70К содержательных строк кода, к тому же ещё использует библиотеку Readline, которая, хоть и поддерживается тем же товарищем, что и Bash, и разрабатывалась в основном для Bash-а, но формально к нему не относится и не имеет кода, специфичного для Bash.

Дальнейшее изложение будет вестись по книге A. Brown, G. Wilson, The Architecture of Open Source Applications и статье J. Garcia et al., Obtaining Ground-Truth Software Architectures. Книга, кстати, весьма годная, редакторы собрали довольно большое количество не очень больших заметок об архитектуре известных проектов с открытым исходным кодом от их авторов или maintainer-ов. С одной стороны, каждый из авторов описывает архитектуру как умеет, поэтому там всё очень неформально, не очень подробно и далеко не всегда соответствует правилам, про которые мы будем рассказывать на этом курсе. С другой стороны, как редакторы справедливо отмечают во введении, архитекторы, которые проектируют здания, в ходе учёбы изучают сотни проектов зданий и критики этих проектов от профессионалов, тогда как архитекторы ПО, как правило, знакомы только с несколькими крупными системами, и то большинство из них они сами же и проектировали. Связано это с очень большой сложностью ПО и невозможностью зачастую восстановить архитектуру по программе, так что возможность посмотреть, как другие проектируют ПО и поучиться на их ошибках или перенять их хорошие идеи может быть очень ценной.

Вот как выглядит диаграмма, характеризующая на высоком уровне архитектуру Bash (примерно то, что я хотел увидеть на доске на первой паре, наверное):

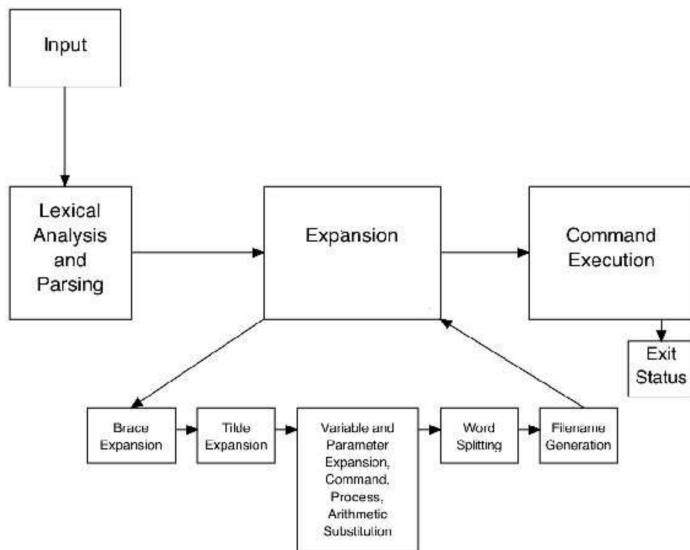


Диаграмма показывает поток данных между основными компонентами системы (в основном, Exit Status на самом деле компонентом системы не является, но я говорил, что неформальные диаграммы из квадратиков и стрелочек очень популярны и весьма полезны). Ввод (с консоли или из файла) поступает на вход лексическому и синтаксическому анализатору, далее последовательность слов, полученная парсером, отдаётся expansion-у, который на самом деле представляет собой конвейер, последовательно применяющий преобразования к последовательности слов. Сначала выполняется подстановка фигурных скобок, затем тильды, затем доллара, затем разделение на слова (после парсера, такие дела), затем раскрытие шаблонов. Дальше то, что получилось, подаётся на вход исполнялке команды, которая отвечает за перенаправление ввода-вывода, пайпы, группы процессов и т.д., в итоге получается результат выполнения команды в виде кода возврата.

Всё общение между компонентами выполняется с помощью структуры WORD_DESC и различных контейнеров, содержащих эти структуры. Вот определение этой структуры:

```

typedef struct word_desc {
    char *word; /* Zero terminated string. */
    int flags; /* Flags associated with this word. */
} WORD_DESC;
  
```

А вот так, например, представляются аргументы команды:

```

typedef struct word_list {
    struct word_list *next;
    WORD_DESC *word;
} WORD_LIST;
  
```

2.1. Ввод с консоли

За ввод с консоли отвечает библиотека Readline, которая отвечает за редактирование командной строки и за хранение истории команд. Она устроена как цикл “читать сим-

вол с клавиатуры - найти команду, ему соответствующую - исполнить её - показать результаты”. Символ считывается как 8-битный символ (в те времена, когда это писалось, юникода ещё не было) и используется как индекс в “таблице диспетчеризации”. В этой таблице может быть либо команда (например, “перейти в начало строки”), либо другая такая же таблица (это для поддержки сочетаний из нескольких символов), либо команда “вывести считанный символ”. Ещё есть макросы (которые просто вставляют во входной поток символы). Все выводимые символы хранятся в буфере редактирования, а когда надо вывести результат на экран, Readline хитро рассчитывает минимальный набор команд управления курсором, который преобразует текущую отображаемую на экране строку в желаемую. Все внутренние данные хранятся в виде 8-битных символов, но Readline знает (теперь) про юникод и умеет его корректно отображать и корректно считать позиции для многобайтовых символов.

Readline может быть расширена добавлением произвольных функций в таблицу диспетчеризации, и Bash этим пользуется, добавляя более 30 своих команд (например, автодополнение).

2.2. Синтаксический разбор

Readline возвращает просто строку, введённую пользователем. Первое, что с ней делается — лексический анализ, то есть, в случае с Bash-ем, разделение по словам и их идентификация. С последним дела обстоят довольно плохо, потому что смысл последовательности символов сильно зависит от контекста, так что лексер и парсер должны тесно общаться друг с другом, чтобы разбирать, например, вот такой ужас:

```
for for in for; do for=for; done; echo $for
```

Эта команда, кстати, вполне корректна и выведет на экран “for”.

Bash — один из немногих шеллов, написанный на lex + bison, о чём, впрочем, автор несколько сожалеет, говоря, что ручная реализация рекурсивным спуском сильно упростила бы дело. Оригинальная грамматика шелла Борна, которую Bash пытается поддерживать, никому не известна, есть грамматика (контекстно-зависимая), стандартизованная POSIX, её-то и реализует Bash (так что грамматика Bash таки известна и документирована).

Интересно, что подстановка alias-ов выполняется лексером. Правда, для этого парсер сообщает ему, разрешена ли в данный момент подстановка. Лексер же отвечает за кавычки и бэкслеш.

Дополнительные проблемы создаёт подстановка результата выполнения команды и программируемое автодополнение, где тоже могут выполняться произвольные команды прямо в процессе разбора другой команды. Для поддержки таких вещей парсер умеет сохранять своё состояние и корректно восстанавливать его после разбора и исполнения “подкоманды”.

Результат работы парсера — одна команда (которая в случае сложных копанд типа for может содержать подкоманды), которая передаётся модулю, отвечающему за подстановки.

2.3. Подстановки

Подстановки (expansions) работают на уровне слов и могут порождать новые слова и списки слов. Они могут быть довольно сложными, например,

```
${parameter:-word}
```

раскрывается в *parameter*, если он установлен, и в *word*, если нет. А

```
pre{one,two,three}post
```

раскрывается в

```
preonepost pretwopost prethreepost
```

Ещё бывает подстановка тильды и арифметическая подстановка. Все они выполняются по порядку, то есть подстановщики организованы во что-то вроде конвейера.

Результат подстановки снова разбирается на слова (при этом это делает код, видимо, отличный от лексера). После разбиения происходит замена шаблонов — каждое слово интерпретируется как потенциальный шаблон и пытается сопоставиться с файлом в файловой системе.

2.4. Исполнение команд

Команды бывают встроенными (которые модифицируют состояние самого шелла, например, `cd`) и внешними, которые сами отдельные программы (например, `cat`). Ещё бывают сложные команды — `if`, `for` и т.д.

Каждая команда позволяет перенаправлять свой ввод и вывод (да, даже в `for` можно направить входной поток из пайпа). Самое сложное в реализации перенаправления — это следить за тем, когда его нужно отменить. Встроенные и внешние команды работают с точки зрения пользователя одинаково, поэтому наивная реализация перенаправления вывода встроенной команды перенаправила бы вывод всего шелла. При этом Bash ещё и следит за файловыми дескрипторами, которые участвуют в редиректе, создавая новые или используя старые при необходимости.

Встроенные команды реализованы как сишные функции, принимающие список слов как аргументы, и работают как “обычные” команды, только без запуска отдельного процесса. Единственная тонкость в том, что некоторые команды принимают как аргумент присваивания (например, `export`), они обрабатывают присваивание по-особому (для этого используются флаги в `WORD_DESC`). Обычные присваивания (которые не в `export` или `declare`) реализованы на самом деле тоже как команды, но парсятся и обрабатываются немного по-особому. Если присваивание стоит перед обычной командой, то оно передаётся команде и действует до её завершения, если присваивание одно на строке, оно действует на весь шелл.

Внешние команды перед запуском ищутся в файловой системе, при этом шелл смотрит на `PATH`. Но если в имени команды есть слеш, то не ищет, а исполняет как есть. При этом единожды найдённая команда запоминается в хеш-таблице и дальше сначала ищется в ней (перехитрить Bash вроде не получится, если он по запомненному пути не нашёл, то пойдёт искать в `PATH`-е). Если команда не нашлась, Bash вызывает функцию, которую можно переопределить, и многие дистрибутивы это делают, предлагая поставить нужный пакет с командой.

Ещё есть некоторые хитрости с управлением процессами, в которых исполняются команды. Есть режим `foreground`, в котором шелл ждёт завершения процесса с командой,

есть background, где шелл запускает команду и тут же читает следующую. При этом Bash умеет переводить команду из одного режима в другой, для чего там есть понятие “Job”, как группа процессов, исполняющая команду. Например, пайпы собирают несколько процессов в один Job, который может быть отправлен в фон или в foreground целиком.

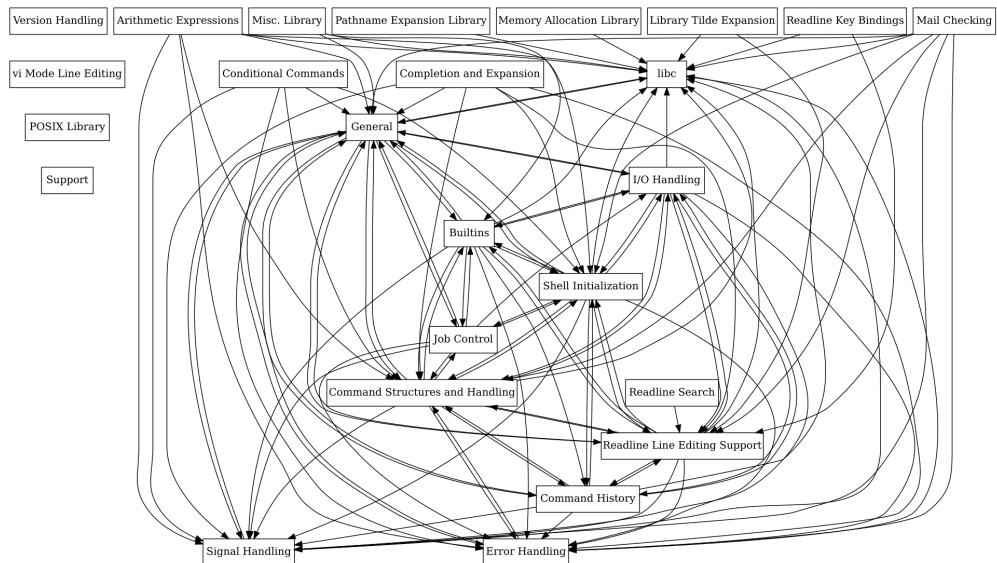
2.5. Lessons learned

Вот кратко вещи, на которые разработчик Bash Chet Ramey обратил внимание в ходе разработки.

- Хорошие комментарии к коммитам со ссылками на багрепорты с шагами воспроизведения сильно облегчают жизнь.
- Хороший набор тестов — тоже, Bash имеет тысячи тестов, покрывающие практически всю неинтерактивную функциональность.
- Сильно помогли стандарты, как внешние на функциональность шелла, так и внутренние на код.
- Хорошая пользовательская документация важна.
- Переиспользование сильно экономит время.

2.6. Как обстоят дела на самом деле

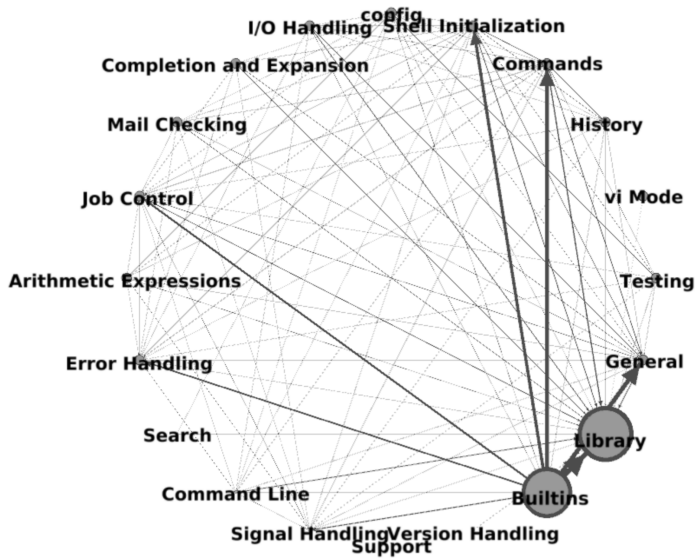
Товарищи из университета Южной Калифорнии исследовали “настоящую” архитектуру Bash с целью получить “базовую” архитектуру, по которой можно было бы оценивать эффективность различных инструментов автоматического восстановления архитектуры. Один аспирант 80 часов копался в исходниках, после чего отправил результаты автору и тот ещё высказал свои соображения. В итоге получилась вот такая структура системы:



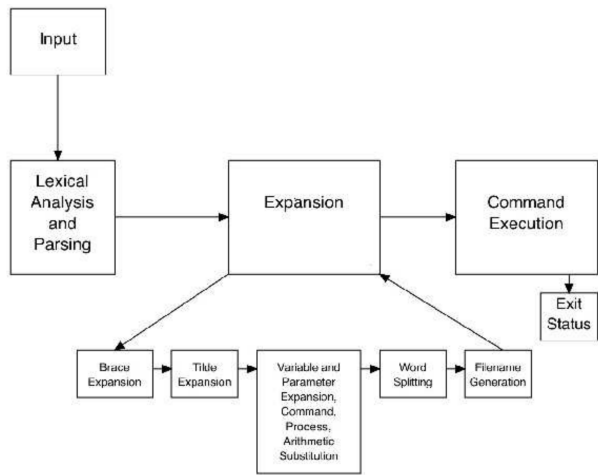
Видно, что зависимостей в коде больше, чем на картинке с концептуальной архитектурой и поток данных на структурной диаграмме совершенно неочевиден. Но некоторые схожести всё-таки есть, например, ввод-вывод реально выделяются в отдельный кластер компонентов.

Bash имеет размер порядка 70К строк кода, около 200 отдельных файлов. В ходе восстановления архитектуры было выделено 25 компонентов, из которых 16 относятся к ядру функциональности системы, 9 — вспомогательные. Выяснилось, что структура папок практически не соответствует компонентам, только два компонента имеют свои отдельные папки в исходниках.

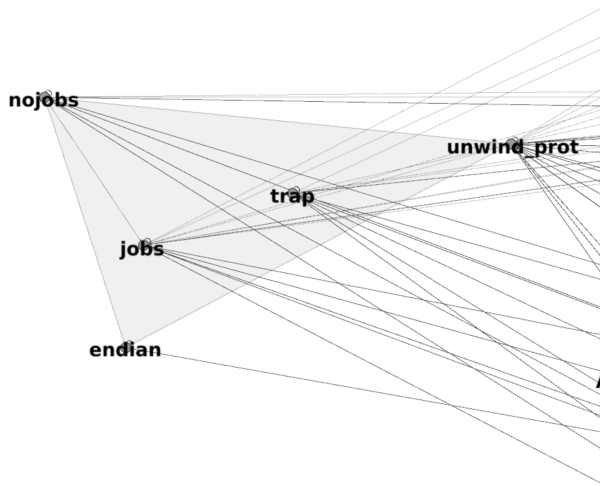
Вот автоматически восстановленная по исходникам архитектура системы, с компонентами и зависимостями:



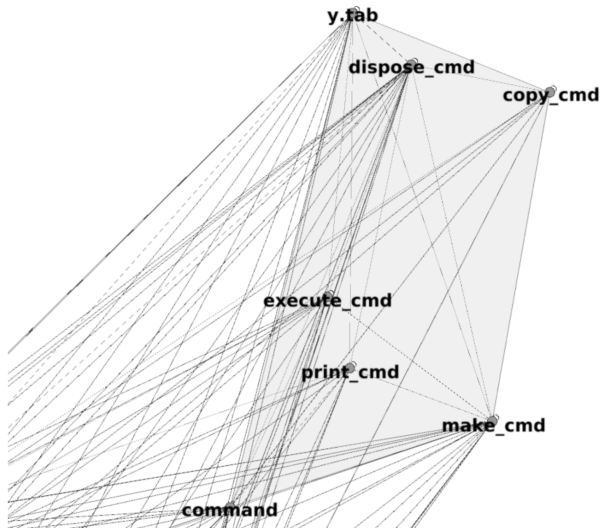
Сравните с исходной концептуальной архитектурой:



Это должно подвести к важной мысли касательно архитектур реальных приложений — есть архитектура как она проектировалась (prescriptive architecture), есть архитектура как она реализована в коде приложения (descriptive architecture), и в ходе эволюции приложения эти архитектуры становятся всё больше и больше непохожими друг на друга. Эти расхождения связаны с понятиями “architectural drift” (привнесение в архитектуру вещей, которых в исходной архитектуре не было, без нарушения принципов исходной архитектуры) и “architectural erosion” (привнесение в реализацию нарушений принципов исходной архитектуры). Для долгоживущих систем архитектурная эрозия становится довольно критичным фактором, приводящим к тому, что исходное разбиение на компоненты перестаёт быть валидным вообще. Bash в этом плане довольно показателен, поскольку ему много лет. Вот увеличенный вид компоненты управления Job-ами:



Как видим, каждая сущность с этой компоненте больше связана с внешними сущностями, чем с сущностями внутри компоненты, то есть coupling очень высок, а cohesion, судя по всему, очень низок. С командами дела обстоят ещё хуже:



Собственно, поэтому важна архитектурная документация и нелишнее наличие архитектора, который следил бы за тем, чтобы код и документация не очень расходились. Иначе приложение быстро превратится в гигантский ком кода, где всё зависит от всего и ломается при любом изменении.

3. Grep

Следующая домашняя задача — реализовать в рамках существующего CLI (ну, у большинства — пока потенциально существующего) команду грег, которая бы искала подстроки в файле или входном потоке. Весь синтаксис грег-а поддерживать не надо, надо уметь принимать регулярные выражения (так, как они реализованы в вашем любимом языке программирования, писать свой парсер регэкспов не нужно) и поддерживать ключи:

- `-i` — нечувствительность к регистру;
- `-w` — поиск только слов целиком;
- `-A n` — распечатать `n` строк после строки с совпадением.

Вот некоторое количество примеров вызова того, что должно получиться:

```
> grep plugin build.gradle
  apply plugin: 'java'
  apply plugin: 'idea'
> cat build.gradle | grep plugin
  apply plugin: 'java'
  apply plugin: 'idea'
> grep -A 2 plugin build.gradle
  apply plugin: 'java'
```

```
apply plugin: 'idea'  
group = 'ru.example'  
version = '1.0'
```

С архитектурной точки зрения эта задача интересна тем, что писать разбор аргументов командной строки вручную — это дело муторное и, кажется, широко распространённое (существуют тысячи консольных утилит, они ведь должны как-то парсить свои аргументы). Поэтому, как обычно, перед тем, как кидаться что-то кодить, надо поискать уже готовые опенсорсные решения. Собственно, это одна из самых важных задач архитектора — во-первых, чувствовать, где можно применить третьесторонние компоненты, а где проще реализовать самим, во-вторых, уметь совершать обоснованный выбор третьейсторонней компоненты среди аналогов, в-третьих, знать кучу всего, чтобы иметь возможность сказать «ага, а вот есть такая штука, она делает то, что нам надо! (и у неё наверняка есть аналоги, потому что я слышал о ней уже три года назад, она наверняка протухла)».

Каких-то неочевидных практик поиска третьесторонних компонент, наверное, не существует, вот немножко очевидных рекомендаций, на всякий случай.

- Миллион леммингов не может ошибаться, поэтому у хорошей либы наверняка много пользователей (правда, тут следует брать поправку на год появления), быстрое гугление покажет, насколько активно сообщество на GitHub, StackOverflow и подобных местах.
- Даже если компонента очень хороша, маленькое сообщество приведёт к проблемам — негде будет найти предыдущий опыт, труднее будет получить совет, если что-то пойдёт не так.
- Если есть несколько фаворитов, имеет смысл потратить время на то, чтобы поковыряться с ними со всеми и выбрать то, что больше подходит или даже просто больше понравилось. Пробовать прежде всего стоит то, что кажется сложным, потому что часто бывает, что что-то простое делается просто, а сложное — вообще никак, и придётся отказаться от выбранной компоненты посреди процесса интеграции.
- Обращайте внимание на лицензию. Даже лицензия типа GPLv3 может быть showstopper-ом для использования идеальной во всех остальных отношениях компоненты. Ни один нормальный архитектор не интегрирует в свою систему компоненту, которая хотя бы потенциально может иметь проблемы с авторскими правами.
 - Поэтому если вы всё ещё выкладываете код на GitHub без явного указания лицензии, подумайте, достаточно ли добродетельный вы человек. Код без лицензии по большинству законов использовать может только автор, даже если он лежит в открытом доступе.

Хочется, чтобы на этом простом примере вы попробовали сделать обоснованный выбор библиотеки разбора параметров командной строки. Выбор может быть не очевиден, но сами по себе такие библиотеки не очень сложные, так что тут реально хорошо подходит тактика «попробовать все и выбрать то, что пришлось по душе больше». Тем не менее, мы «играем во взрослые проекты», поэтому хочется в домашнем задании текстовый документ с описанием того, какие библиотеки были рассмотрены, их достоинства и недостатки и

почему была выбрана именно та библиотека, которую вы в итоге выберите. Аргумент «Я всю жизнь ей пользовался» считается в данном случае невалидным.

Задачу надо сдавать в отдельной ветке, отведённой от ветки CLI, дедлайна у неё нет, но помните, что все задачи надо обязательно сдать до зачёта.

А сейчас мы продолжим то, на чём остановились на прошлой паре — проектировать CLI. Можно выйти к доске, нарисовать пару диаграмм и совместными усилиями прийти к “идеальной” архитектуре CLI.