SPIS TREŚCI

SPIS RYSUNKÓW

SPIS LISTINGÓW

SPIS TABEL

1. INTRODUCTION

What is the problem and why is it worth solving?

Evolutionary methods efficiency optimization using distributed systems

3. EXISTING SOLUTIONS OVERVIEW

4. WHAT IS MY SOLUTION TO A GIVEN PROBLEM?

4.1. IDEA

The main goal of this theses is to create a tool, which allows to integrate a high-performance implementation of the genetic algorithm into user application and use it for the arbitrary optimization problem with minimum changes required to the existing codebase. Additionally, this should be a convenient toolbox for theoretical and practical problems which might be resolved using genetic algorithms, providing a ready-made implementation of the most popular algorithms with extra features, which save users from implementing one themselves.

4.2. KEY CONCEPTS

As mentioned previously, there already are genetic algorithms implementations available in open-source, which may satisfy some needs, but still offer too small to be universal solution and base on the architectures, which don't enable some crucial enhancements. For the tool to be better than existing ones, it has to be built upon the concepts which meet the needs of a broader range of scientists and developers, as well as provide more comfortable user experience, by reducing boilerplate, lowering entry-level and grant united environment.

1. **Arbitrary representation of individual**
   There is a wide set of problems which might be resolved using genetic algorithms. Still, this optimization technique is rarely used in business applications and mainly considered as a theoretical solution. One of the reasons for such phenomenon is the difficulty which comes with a try to use genetic algorithm optimization without a previous plan to do so. The process of reshaping the existing codebase in such way, so typical genetic algorithm could be applying often demands big costs and high level of understanding GA workflow and its possibilities.
   Depending on the problem a user is trying to solve, a standard representation of each candidate solution would be an array of bits, integers or real numbers. The main property that makes these genetic representations convenient is that their parts are easily aligned due to their fixed size, which facilitates simple crossover operations. For the purposes of genetic programming and evolutionary programming tree-like and graph-form representation are used respectively, which causes the necessity in different implementations for each problem.

A solution to this problem is an implementation of the genetic algorithm, which is able to work on generic types, which represents a candidate solution for the applied problem. Due to this property, it is possible to reuse the implementation for different problems and create a generic tool, which is decoupled from the problem it is applied to.

2. **Local and distributed parallelization**
   Genetic algorithms are generally known as a time-consuming technique, as it is often used with complex, high-dimensional problems. Evolution cycle may take up to hours and days of continuous computing, depending on the parameters of the algorithm and hardware in use, which comes with a great cost and very low flexibility, as with every mistake or change was done to algorithm implementation, evolution needs to be re-executed. On the other hand, even small upgrades in speed may result as a major cost cut.
   For this reason, it is highly important to use the maximum of the given resources, avoiding unreasonable workloads. It may be achieved through the computation parallelization between multiple virtual CPU cores within a single machine and distributing computation in a cluster of working nodes. For the better result, two approaches may be combined and used together.

3. **Large size of the population**
   The important parameter of the genetic algorithm is population size. It controls how many individuals are going to be processed by algorithm during one evolution step. When applying to a problem with a wide range of possible solutions the *algorithm tends to find the optimal solution faster, while operating on a large number of candidates at once*. It makes every population more diverse, which leads to faster identification of better candidate and also allows to cover more subsets of function arguments at the beginning of the evolution process.
   A typical limitation of the number of objects one can use at a time is the capacity of RAM space. Working with large sets of data, while keeping them in memory like collections in Java or any other programming languages, may lead to exceeding the capabilities of the hardware, resulting in a crash of application or undefined behavior. In order not to limit the users, in their options it must be possible to store the parts or whole population on disk (which usually is more spacious than RAM), while processing the population in parts, which may be safely stored in faster RAM. The choice of the way, how the algorithm will store its intermediate data should be open to the end users, enabling them to make a decision upon their own preferences, which may be connected to the existing software and hardware. These parameters will be different among different applications the tool is used for, so both options, as well as their combination, have to

available “out of the box”.

4. **On-demand evolution**
   Genetic algorithm optimization is generally used for finding the best solution for a defined problem, which is the fittest candidate to the provided fitness function. It means, that the best solution is defined relative to all the solutions assigned earlier and cannot be proved as the optimal one, even when it is. But when applying to a nontrivial problem, finding the optimal solution is not always cost-effective, as it may take up to exponential time. For this reason, it is common to compute a ’good enough’ solution in reasonable time. Depending on the business reasons, this time varies so sometimes it is possible to wait until the better solution is computed and sometimes it is not. For instance, if the application which uses genetic algorithm has a real-time interface and its user is interested in gaining a solution for the particular case in the range of seconds, it is justifiable to provide a solution which may not be optimal but is close to it.
   On the other hand, when a user is ready to allow the application to take its time in order to provide a better result, it has to still be possible to resume the evolution. Again the application should not limit users in such important cases, as the end user needs may vary depending on different circumstances, which may not be predicted in advance. The evolution of a solution to a given problem should be available on-demand, without forcing unreasonable resource uses, and adjust to a specific case as much as possible.

5. **Flexible evolution flow**
   Genetic algorithm workflow depends on many parameters as well as provided genetic operators (selection, crossover, mutation). Each one of them decides how the computation will be processed, for how long, how many values will be tested, how the solution candidates will be compared between each other etc. These parameters should be wisely chosen, according to the specification of the problem, which genetic algorithm is applied to. But even within the same environment, the set of settings should be changed multiple times in order to achieve the best result in the shortest time frame. Improving the fitness function during the evolution process and adjusting the size of a population depending on the average number of unque individuals may serve as examples of self-improving performed by a genetic algorithm on the fly. Such kind of real-time modifications is often an advanced part of the implementation, which demands a deep understanding of the domain. As a workaround, sometimes the same effect is attempted to be achieved by a combination of multiple genetic algorithms, each with a different configuration, combined in a pipeline, where the result of a predecessor serves as a starting point for a successor. Still, this solution rarely allows achieving the same result, as each stage of the pipeline serves as a standalone processing workflow.
   The ability to change its configuration during the process is complicated, but still an

important feature, which enables to use GA implementation for complex practical problems.

4.3. IMPLEMENTATION OVERVIEW

The solution is presented as an open-source library, that provides the major part of genetic algorithm implementation, which can be shared between different application. Base algorithm consist of a pipeline separated by three stages: selection, crossover and mutation, which evolves the population of individuals, increasing average level of adjustment and finding the best individual according to provided fitness function; preimplemented most popular strategies of genetic selection, crossover and mutation, with possibility of extanding with the additional ones for specific reasons and a set of platforms which allow to run the pipeline in a different ways, best suited for existing hardware. The library is supplied with key features, basing on the concepts described earlier, wrapping all the components into sufficient framework.

4.3.1. Technology stack

Scala programming language

Scala is a modern multi-paradigm programming language designed to express common programming patterns in a concise, elegant, and type-safe way. Its name stands for "scalable language." The language is so named because it was designed to grow with the demands of its users. It can be applied to a wide range of programming tasks, from writing small scripts to building large systems. Scala is easy to get into. It runs on the standard Java platform and interoperates seamlessly with all Java libraries. It's quite a good language for writing scripts that pull together Java components. But it can apply its strengths even more when used for building large systems and frameworks of reusable components. [2]

Technically, Scala is a blend of object-oriented and functional programming concepts in a statically typed language. The fusion of object-oriented and functional programming shows up in many different aspects of Scala; it is probably more pervasive than in any other widely used language. The two programming styles have complementary strengths when it comes to scalability. Scala's functional programming constructs make it easy to build interesting things quickly from simple parts. Its object-oriented constructs make it easy to structure larger systems and adapt them to new demands. The combination of both styles in ScaalMatei Zaharia, "Spark: Cluster Computing with Working Sets makes it possible to express new kinds of programming patterns and component abstractions. It also leads to a legible and concise programming style [2]

Due to its programming style and variety of features Scala serves a great tool for developing this library. A combination of paradigms is extremely usefull in this case and

was crutial during selection of programming language. Both of the programming styles were heavily used in order to gain the best possible result:

- Object-oriented style
  Scala is an object-oriented language in pure form: every value is an object and every operation is a method call. [2] It provides modularity and maintainability: all application is split into a self-sufficient containers, which store the relevant data and provide available operations, that may be performed on its data. These containers may be used as the assembling parts for another container, passed as function parameter and returned as function result. Such approach allows to keep track of the growing codebase, alligning its components into domain hierarchies and letting its parts to comunicate with one another.
- Functional style
  Functional programming provides expressiveness, composability and conciseness. Building the library with mathematical-like functions makes it more understandable, replacing the mutable state from the application. As the result, it becomes clear, unequivocal and may be safely used in multi-thread environment. One of the most important assets, which comes from functional programming, is scalability. A well constructed application may easily be enlarged for user needs to handle a growing amount of work; it is more adaptive to the changing needs and the performance of such system improves proportionally to the additional hardware.

Parallel collections

Encouraging usage of immutable objects, from version 2.8 Scala provides new collection API, which contains parallel collections package as part of it. Parallel collections were included in the Scala standard library in an effort to facilitate parallel programming by sparing users from low-level parallelization details, meanwhile providing them with a familiar and simple high-level abstraction. The idea is simple– collections are a well-understood and frequently-used programming abstraction. And given their regularity, they're able to be efficiently parallelized, transparently. By allowing a user to "swap out" sequential collections for ones that are operated on in parallel, Scala's parallel collections take a large step forward in enabling parallelism to be easily brought into more code. [3]

How parallel collections are built? What may be acquired with scala parallel collections?

Apache Spark

Apache Spark is a high-performance, general-purpose distributed computing system that has become the most active Apache open source project, with more than 1,000 active contributors. Spark enables users to process large quantities of data, beyond what can fit on a single machine, with a high-level, relatively easy-to-use API. Spark's design and interface are unique, and it is one of the fastest systems of its kind. Uniquely, Spark allows

to write the logic of data transformations and machine learning algorithms in a way that is parallelizable, but relatively system agnostic. So it is often possible to write computations that are fast for distributed storage systems of varying kind and size.

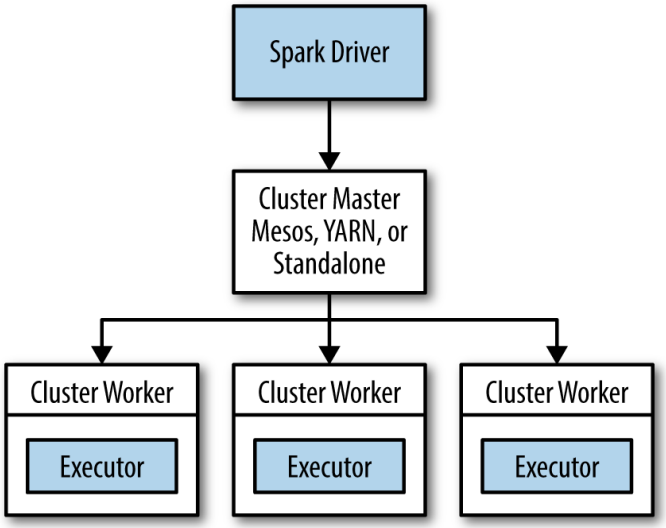
On the generality side, Spark is designed to cover a wide range of workloads that previously required separate distributed systems, including batch applications, iterative algorithms, interactive queries, and streaming. By supporting these workloads in the same engine, Spark makes it easy and inexpensive to combine different processing types, which is often necessary in production data analysis pipelines.[1]

Spark can run over a variety of cluster managers to access the machines in a cluster. The easiest way is to run Spark by iteself on a set of machines. For this purpose Spark comes with built-in Standalone mode. Spark's Standalone manager offers a simple way to run applications on a cluster. It consists of a master and multiple workers, each with a configured amount of memory and CPU cores. When application is submited, one can choose how much memory its executors will use, as well as the total number of cores across all executors.

Spark Core contains the basic functionality of Spark, including components for task scheduling, memory management, fault recovery, interacting with storage systems, and more. Spark Core is also home to the API that defines resilient distributed data-sets (RDDs), which are Spark's main programming abstraction. RDDs represent a collection of items distributed across many compute nodes that can be manipulated in parallel. Spark Core provides many APIs for building and manipulating these collections. RDDs offer two types of operations: transformations and actions. Transformations construct a new RDD from a previous one. For example, one common transformation is filtering data that matches a predicate. Actions are operations, which produce different result than the transformation, like collecting the elements into a local collection on master node or counting the number of elements in RDD. Transformations and actions are different because of the way Spark computes RDDs. Although new RDDs can be defined in any time, Spark computes them only in a lazy fashion — that is, the first time they are used in an action. The transformations, on the other hand, are scheduled to a DAG (directed acyclic graph) of computations. This approach has a lot of advantages, among them: effective fault tollerance and ability to make a lot of optimization decisions before actually running operations. This would not be possible if it executed computations as soon as it got it.

In distributed mode, Spark uses a master/slave architecture with one central coordinator and many distributed workers. The central coordinator is called the driver. The driver communicates with a potentially large number of distributed workers called executors. The driver runs in its own Java process and each executor is a separate Java process. A driver and its executors are together termed a Spark application. Then a Spark application is launched on a set of machines using early described cluster manager.

This architectures allows to distribute the data between different executor nodes and



Pic. 4.1. The components of a distributed Spark application [1].

Level	Space used	CPU time	In memory	On disk	Comments
MEMORY_ONLY	High	Low	Y	N	
MEMORY_ONLY_SER	Low	High	Y	N	
MEMORY_AND_DISK	High	Medium	Some	Some	Spills to disk if there is too much data to fit in memory.
MEMORY_AND_DISK_SER	Low	High	Some	Some	Spills to disk if there is too much data to fit in memory. Stores serialized representation in memory.
DISK_ONLY	Low	High	N	Y	

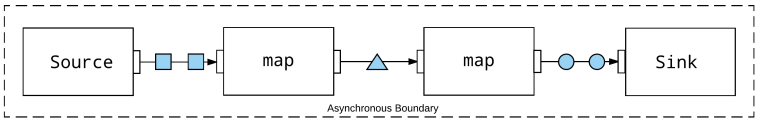
Pic. 4.2. Persistence levels from org.apache.spark.storage.StorageLevel and pyspark.StorageLevel [1].

perform operations on it locally until the shuffling is needed. Only then the data is shared between different nodes and the network communication is handled. In a distributed program, communication is very expensive, so laying out data to minimize network traffic can greatly improve performance. Much like how a single-node program needs to choose the right data structure for a collection of records, Spark programs can choose to control their RDDs' partitioning to reduce communication.

In order to be able to transforms large sets of data without redundant recomputation Spark provides data persistence. When the Spark is asked to persist an RDD, the nodes that compute the RDD store their partitions. Spark has many levels of persistence to choose from based on what the goals are. The list of storage levels with performance comparison may is shown on 4.2. Combining different persistence modes, an executor can effectively share the data between its disk and RAM.

Akka Streams

Akka Streams is a library to process and transfer a sequence of elements using bounded buffer space. This latter property is what is ususally referred as boundedness and it is the



Pic. 4.3. A typical stream flow

defining feature of Akka Streams. Translated to everyday terms it is possible to express a chain of processing entities, each executing independently (and possibly concurrently) from the others while only buffering a limited number of elements at any given time.

ToDo: more info

4.3.2. Core components

ToDo: add introduction

- Individual
  An individual is a core component of genetic algorithm, often also reffered to as a *genotype, chromosome* or *a candidate*, which represents a candidate solution to a given problem. The number of individuals are evolved during the algorithm workflow, mixed and mutated in order to find the fittest one according to provided fitness function. This concept is purely abstract in the library, encoded as a generic type which may have an arbitrary representation defined by the users depending on their needs.
  In order to handle an individual through the process of evolution there should be defined a set of operators, where each encapsulates logic of some part of evolution procedure.

  · Fitness
    Fitness is a type class which represents a function of type `I => Double`, where `I` is type of individual. This function summarises how close a given design solution is to achieving the set aims. A ftness function must be devised for each problem to be solved, thus should be provided by the user and may not be preimplemented. Fitness function value is main criterion of future selection and is very important attribute of individual representation.
    ToDo - how much does it cost

  · Join
    Join is a type class, which represents a cross function of type `(I, I) => (I, I)`, where `I` is type of individual. This function defines how two instances of type `I` may

be combined together to produce a new pair of `I`. It describes a crossover operation, due to which all the individuals from a population combined into pairs will be mixed during evolution process. An instance of the type class may be created from a single function of type `(I, I) => (I, I)`, as well as a function of type `(I, I) => I`. The latter case is suitable when identical function is used to compute both of result elements or in case of commutative operation. These cases were separated in order to enable memory allocation optimization by reusing the same product object to construct result pair.

· Modification

Modification is a type class, which represents a mutation function of type `I => I`, where `I` is type of individual. This operator is used to maintain generic diversity from one generation of a population to another. It is analogous to biological mutation. Mutation alters one or more gene values in a chromosome from its initial state, during which the solution may change entirely from the former shape. Because of its loose representation (even identity function may be considered as modification) it is important to provide a set of laws, which will distinguish a modification, which actually brings a diversity into the population. A *lawfull* modification function is one, which holds following properties:

1. A modified instance does not equal to original one (*e.i modification function does not equal to identity function*)

```
modify(i) != i
```

2. After a certain number of modification the same input produces different outputs (*e.i. modification function is randomized or depends on outer variables*)

```
def modify5(i: I) = modify(modify(modify(modify(modify(i)))))
modify5(i) != modify5(i)
```

· Scheme

Scheme is a type class, which is responsible for a creation of new individual. This is additional function, which may be replaced by predefined initial population during the start of evolution process, but still offers more functionality and may be used when individuals are created randomly, from limitless iteratator, smaller collection, which should be cycled or even single instance of individual. This concept is implemented as a covariant class, so the scheme object of parent class may be replaced by the scheme object of child class. It allows to easily use this class with hierarchies, avoiding redundant code.

- Population
  Population is a collecion of individuals, which holds them during single evolution step. Individuals are processed together and are viewed only as a part of population from the perspective of algorithm. During the evolution step the representation of population or holding type may be changed (e.g. after selection stage further processed population contains pair of individuals, which are meant to be mixed later).
  From implementation perspective, population is a type alias to Scala vector collection. The choise of data structure was motivated by fast average performance and effectively constant random access time in particular.

- Evolution environment
  Evolution Environment is a component, which taking the set of options constructs endless evolution flow of given population and makes it ready to be run. It is a parent type to all implementations of different approaches in processing computations, which may be sequential or parallel, local or distributed. This concept is very important from the usage perspective as it devides the implementation of genetic algorithm from its usage, so the end user may even not be aware of the way how the evolution is performed. It is also the end point of all the dependent components, as there is no direct dependency to any of concrete implementation of Evolution Environment. This abstraction may be considered as high level representation of genetic algorithm itself. The list of concrete derivatives of this interace contains:
  - Local environment
    This type of environment presumes local evoluation of evolution flow. It is the richest environments in terms of possible configurations. The logic, behind the strategy of applying genetic operators and evaluating fitness values is decoupled from one another into instances of `EvolutionCompanion` and `Fitness Evaluator` classes respectively. For each of the strategies the are sequential and parallel versions, with additional fitness caching option for repetitive individuals within population. Any fitness evaluation strategy may combined with any possible technique of applying genetic operators, creating a number of possibilites ready for use.

  - Asynchronous environment
    Asynchronous environment delegates the function of fitness evaluation to another system, by sending non-blocking asynchronous calls for every element of the population. It may be extremely usefull for the problems with heavy fitness functions. This approach allows to compute fitness values on different machine or even cluster of machines, as well as using different low-level programming language or some specific optimization, with no need to switch the entire application to another platform. This type of environment also fits perfectly to cloud computing solutions, as fitness evaluation is entirely isolated from the rest of application. However, applying

genetic operators in same way needs to send a lot more data over the network, which negates the benefits one can gain from this approach. For this reason, genetic operators are applied locally with any of the strategies described earlier.

  - Distributed environment
    Distributed environment is very different from the previous options. Here the evolution is performed on the population in terms Spark RDD, distributed over the cluster. This technique enables processing of very large population, since it is split between many executor nodes. Once the evolution step is done executors shuffle the data between each other in order to perform selection of individuals for the next population and continue working on the local values. It minimizes the network communication, which may be a bottle-neck in such applications.

- Evolution flow *ToDo - add description*
- Genetic operators *ToDo - add description*
- Fitness caching *ToDo - add description*

4.4. USAGE

The library artifact may be added into the project as dependency using maven tool or sbt. After this, if library requirements are met, imported classes are ready to be used.

4.4.1. Requirements

- Scala 2.12.6
- sbt 1.2.1
- Apache Spark 2.3

4.4.2. API

ToDo - add info

4.4.3. Preimplemented algorithms

ToDo - add info

4.4.4. Evolution interruption

ToDo - add info

How can I enhance this solution?

BIBLIOGRAFIA

[1] Karau, H., Konwinski, A., Wendell, P., Zaharia, M., *Learning Spark: Lightning-Fast Big Data Analytics*, 1st wyd. (O'Reilly Media, Inc., 2015).

[2] Odersky, M., Spoon, L., Venners, B., *Programming in Scala: Updated for Scala 2.12*, 3rd wyd. (Artima Incorporation, USA, 2016).

[3] Prokopec, A., Miller, H., *Parallel collections overview*, `https://docs.scala-lang.org/overviews/parallel-collections/overview.html`.