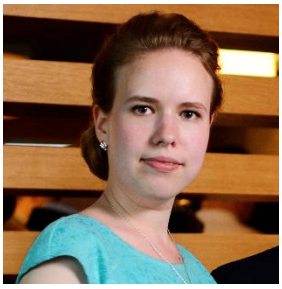
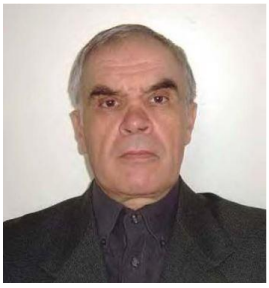
УДК 004.6:37

ВЫБОР ЯЗЫКА ПРОГРАММИРОВАНИЯ ДЛЯ РЕШЕНИЯ ЗАДАЧ, СВЯЗАННЫХ С ПРИМЕНЕНИЕМ ТЕХНОЛОГИИ BIG DATA



В.Д. Львович
Магистрантка кафедры информатики БГУИР



В.Я. Анисимов
Доцент кафедры информатики БГУИР

А.Л. Хотеев
Доцент кафедры информатики БГУИР

М.В. Стержанов
Доцент кафедры информатики БГУИР

Белорусский государственный университет информатики и радиоэлектроники, Республика Беларусь
E-mail: vikalvovich2@gmail.com

В.Д. Львович
С 2018 года обучается в магистратуре на факультете компьютерных систем и сетей Белорусского государственного университета информатики и радиоэлектроники. E-mail: vikalvovich2@gmail.com.

В.Я. Анисимов
Доцент кафедры информатики Белорусского государственного университета информатики и радиоэлектроники, кандидат физико-математических наук. E-mail: anissimov@bsuir.by.

А.Л. Хотеев
Доцент кафедры информатики Белорусского государственного университета информатики и радиоэлектроники, кандидат физико-математических наук. E-mail: al_khot@mail.ru.

М.В. Стержанов
Доцент кафедры информатики Белорусского государственного университета информатики и радиоэлектроники. E-mail: sterjanov@bsuir.by.

Аннотация. Целью работы являлось выявление языков программирования, чаще всего используемых для решения задач, связанных с анализом данных и применением технологии Big Data; описание таких языков и их сравнительный анализ. В результате было определены наиболее популярные языки; выявлены их особенности, позволившие им стать часто используемыми для обработки больших объемов информации; сделаны соответствующие выводы.

Ключевые слова: языки программирования, анализ данных, Big Data, Java, Python, Scala, R.

В современном мире технология Big Data получила широкое распространение. Она нашла свое применение в медицине, торговле, образовании, маркетинге и многих других сферах. И это неслучайно, ведь объем существующих данных с каждым днем увеличивается. Однако такие данные имеют ценность лишь в том случае, если из них можно каким-то образом получить конкретную информацию, применимую для вполне определенных целей.

Потоки больших данных чаще всего автоматически генерируются посредством какой-либо машины и без участия человека, у них отсутствует конкретная структура, а сами по себе они могут даже не представлять особой ценности [1]. В связи с этим для полноценного

анализа больших данных нередко необходимо использовать технологии, отличные от традиционных, усовершенствованные алгоритмы фильтрации, методы прогнозирования и моделирования.

Основой практически любого проекта, связанного с обработкой больших объемов данных, является программирование. Посредством написания некоторого программного кода осуществляется реализация соответствующего алгоритма для решения поставленной задачи. При этом выбор языка, на котором правильнее всего будет осуществить реализацию задуманного решения, зависит от множества факторов: функциональных возможностей языка, поддерживаемых утилит и библиотек, эффективности, масштабируемости и других факторов.

На сегодняшний день к наиболее популярным языкам программирования, применяемым в контексте Big Data, можно отнести Java, Python, Scala и R. Каждый из них имеет некоторые особенности, позволяющие говорить о широкой применимости для обработки больших объемов информации.

К особенностям языка Java, которые позволяют успешно применять его при обработке больших объемов информации, можно отнести:

– отсутствие зависимости от платформы;

– статическую типизацию;

– портативность;

– использование в качестве языка разработки Apache Hadoop, Apache Kafka.

Java считается одним из самых популярных языков программирования. Благодаря данному факту, а также тому, что этот язык программирования является платформонезависимым, он часто применяется и в случае решения задач, связанных с Big Data. При этом существует большое количество библиотек, позволяющих с помощью Java осуществлять интеллектуальный анализ данных. В частности, к таким библиотекам относятся Java Data Mining Packag и Weka.

Можно выделить следующие особенности языка Python, повлиявшие на его популярность:

– простота и универсальность;

– интерпретируемый язык;

– сильная типизация;

– расширяемость и портативность.

Таким образом, использование Python позволяет осуществлять быстрый анализ данных, в том числе и благодаря наличию разнообразных специализированных библиотек, связанных с анализом информации.

В свою очередь, к особенностям языка Scala можно отнести:

– поддержку объектно-ориентированного и функционального программирования;

– взаимодействие с библиотеками Java;

– статическую типизацию;

– проверку типов на этапе компиляции;

– лаконичность и портативность.

Стоит отметить, что Apache Spark изначально написан на языке Scala. Однако на данный момент он осуществляет поддержку и других популярных языков – Java, Python и даже R.

Основными особенностями языка R, которые позволяют ему активно использоваться при работе с Big Data, являются:

– направленность на статистический анализ;

– существование специфических пакетов, применяемых для работы с большими объемами информации;

– широкие возможности для визуализации.

R является не только языком программирования, но и средой для вычислений, работы с данными и графического отображения полученных результатов. Основное применение R заключается в осуществлении работы со статистическими данными. Поэтому считается, что он нацелен на решение лишь ограниченного круга задач, хотя на самом деле обладает определенной универсальностью.

Отличительной особенностью R является то, что помимо широких функциональных возможностей, он позволяет также разнообразными способами представлять и отображать имеющиеся данные. Однако в этом языке программирования достаточно трудно разобраться. Особенно по сравнению с традиционными и более известными языками. А значит, использование R предполагает наличие у программиста определенной специализации.

Большинству программистов комфортнее работать с привычными и распространенными языками программирования. Среди рассматриваемых языков к таким можно отнести Java и Python. При этом Python является более лаконичным, чем Java. То есть число строк программного кода, написанного на Java, как правило, в значительной мере превышает число строк программного кода, написанного на Python и решающего аналогичную задачу. Однако недостатком Python является динамическая типизация, что может привести к возникновению ошибок из-за несоответствия типов, которые не будет изначально обнаружены. Также Java работает гораздо быстрее, чем Python. Для работы с Hadoop можно использовать как Java, так и Python, несмотря на то что Hadoop написан именно на Java.

Таким образом, Python отлично подходит для ситуации, когда необходимо анализировать какие-либо статистические данные и при этом создавать некоторую полноценную интегрированную в производство систему. Стоит отметить, что Python широко используется в случае машинного обучения. Java же применяется, в первую очередь, когда особое значение имеют такие характеристики, как скорость и производительность.

Таким образом, основными языками, предполагающими работу со статистическими данными, являются Python и R. Первый – благодаря наличию специализированных библиотек, а второй – в связи с тем, что он считается языком программирования для статистических вычислений и графического анализа [2].

Scala в скорости практически не уступает Java и значительно превышает Python. При этом язык Scala в основном используется для работы с Apache Spark, и поэтому считается в определенной степени узкоспециализированным. Однако сейчас, когда применение для решения задач Apache Spark вовсе не обязывает писать программный код на Scala, как это было ранее, данный язык перестает быть таким часто используемым. Хотя стоит отметить, что существует мнение, что наиболее эффективную разработку на Spark легче всего проводить на языке программирования Scala [3]. Также нельзя забывать, что данный язык программирования является не только объектно-ориентированным, но и функциональным языком в отличие от Java. При этом программный код на Scala гораздо лаконичнее аналогичного программного кода на Java.

Изначально Scala создавался с целью улучшения Java. В том числе и для упрощения изучения и восприятия начинающими программистами. Однако в результате получилось, что, несмотря на свою видимую простоту, язык Scala является достаточно сложным в освоении. Стоит отметить, что Scala полностью совместим с Java. Поскольку Scala работает на JVM, классы и библиотеки Java могут использоваться непосредственно в программном коде на языке Scala, и наоборот, в программном коде на Java можно применять возможности Scala.

Поскольку хранение данных значительно упростилось, их объем существенно возрос, а структурность перестала иметь большое значение, возникла необходимость разрабатывать специализированные алгоритмы для интеллектуального анализа больших объемов информации. В связи с этим существенное значение приобрел выбор правильного языка программирования, позволяющего упростить решение поставленной задачи. Ведь работа с тем же языком, что использовался программистом ранее, далеко не всегда будет являться верным решением.

При выборе языка программирования, применяемого для работы с большими объемами данных и Big Data, в первую очередь необходимо определиться с целью. Например, язык R хорошо подойдет для анализа данных, но не для работы с «Интернетом вещей». Выявление основных особенностей решаемой задачи позволяет легко определиться с используемыми технологиями и языками программирования. Однако стоит помнить, что иногда подходящих языков программирования может быть несколько. В таком случае достаточно просто разобраться с тем, какие преимущества и недостатки играют наибольшую роль, а какие являются несущественными в рамках решаемой задачи, и выбрать правильный язык программирования.

Литература

[1]. Фрэнкс Б. Укрощение больших данных. Как извлекать знания из массивов информации с помощью глубокой аналитики / пер. с англ. А. Баранова. – М.: Манн, Иванов и Фербер, 2014. – 352 с. УДК 330.47

[2]. Кабаков Р.И. R в действии. Анализ и визуализация данных в программе R / пер. с англ. П.А. Волковой. – М.: ДМК Пресс, 2014. – 588 с. УДК 311:004.9R

[3]. Карау Х., Уоррен Р. Эффективный Spark. Масштабирование и оптимизация. – СПб.: Питер, 2018. – 352 с. УДК 004.62

[4]. Хостманн К. Scala для нетерпеливых. – М.: ДМК Пресс, 2013. – 408 с. УДК 004.432.42Scala

PROGRAMMING LANGUAGE SELECTION FOR SOLVING TASKS RELATED TO BIG DATA

V.D. LVOVICH
Master of the Informatics Department of the BSUIR

V.Y. ANISIMOV
Associate professor of the Informatics Department of the BSUIR

A.L. KHOTEEV
Associate professor of the Informatics Department of the BSUIR

M.V. STERJANOV
Associate professor of the Informatics Department of the BSUIR

Abstract. The purpose was identifying programming languages, most often used to solve problems associated with data analysis and used in big data projects; description the languages and their comparative analysis. As a result, the most popular languages were identified; their specific properties, which allowed them to become frequently used for data analysis, were revealed; appropriate conclusions were made.

Keywords: Programming languages, Data analysis, Big Data, Java, Python, Scala, R.