CMPSC 380
Principles of Database Systems
Fall 2014

Laboratory Assignment Seven
Implementing and Using a Serialization Benchmarking Framework

Introduction

Chapter 23 of your textbook introduces the basics of the eXtensible markup language (XML), explaining how this semi-structured approach to data management has become the defacto way to exchange data. In this laboratory assignment, you will explore some of the trade-offs associated with the use of XML, with a particular focus on the time and space overheads associated with storing data in this format. In particular, you will implement a benchmarking framework that can serialize a Java object to both the default and XML formats. Finally, you and your partner will write a report that describes and explains the performance trends revealed by your benchmark.

Learning About Serialization Methods

The Java programming language provides a default serialization method that transforms a Java object, resident in the heap of the Java virtual machine (JVM), into a format that can be saved to the disk. Why would it be useful to serialize a Java object? You can learn more about how Java performs serialization by studying the following Web site: `http://docs.oracle.com/javase/7/docs/technotes/guides/serialization/index.html`. After learning more about using this default approach to serialization, what do you think are the trade-offs associated with it?

Of course, there are alternatives to Java's standard technique for serialization. For instance, XStream gives a programmer the ability to write a heap-resident object to an XML file. In comparison to using default serialization, what would be the benefits of storing objects in XML files using XStream? You can learn more about how XStream works by visiting this Web site: `http://xstream.codehaus.org/`. You can learn many of the basics about XStream by reading the tutorials that are available on the aforementioned Web site; the "Two Minute Tutorial" provides a simple introduction to the steps that you must take to serialize and deserialize a Java object.

Using the Default and XML-Based Serializers in Java

You and your partner should create a Git repository, using Bitbucket, that you can use to complete this laboratory assignment. To start off this project, one of you should go into the "share" repository for this course and run the "`git pull`" command. After investigating the directories for this assignment, you will find the "`lab7/src/`" and "`lab7/lib/`" directories. What files do these directories contain? How does this program work? Please note that the provided version of the benchmark does not yet include the features necessary to evaluate the performance of default and XML serialization—you and your partner will add this code as part of the laboratory assignment.

Make sure that both team members are able to compile and run the current version of the "`StorageAndRetrievalBenchmark.java`" program. What output does it produce? What is the meaning of this output? You should note that the provided Java archive (JAR) files are older versions of XStream; interested students may investigate the downloading and installation of the version of this library that was released in February of 2014. However, the benchmarks should work correctly with a slightly older version of XStream. Please see the course instructor if you are not able to compile and run this preliminary version of the benchmarks.

Implementing and Using Data Serialization Benchmarks

The Java programming language provides a default serialization method for any Java class that "`implements java.io.Serializable`"—it also requires the use of classes such as `FileOutputStream` and `ObjectOutputStream`. Can you add a default serialization benchmark that will store and retrieve one of the randomly generated objects? How big are the files that this method creates? You should also add an XStream-based benchmark that can store and retrieve the same objects. In comparison to the default serializer, how long does it take to produce the XML-based representation? How big are the XML files? What are the benefits of using XML for serialization?

After finishing the implementation of your benchmarks, you and your partner should conduct experiments to compare and contrast the performance of the default and XML-based serialization and deserialization methods. Your experiments should vary the number and content (e.g., both character and numerical values) of tree nodes in the randomly generated object. You should also measure the time overhead associated with storing and retrieving the object and the space overhead of the serialized files. Finally, you must write a report that analyzes and explains your results.

Summary of the Required Deliverables

You and your partner should always use a Git repository, hosted by Bitbucket, to store the source code, XML files, benchmark output, and all of the other deliverables required by this assignment. The repository must be shared with the course instructor and the version control log should accurately reflect each student's contribution to the assignment. In addition, this assignment invites your partnership to submit one printed version of the following deliverables; each member should write and submit their own version of the first deliverable. Please see the instructor if you have questions about the deliverables that you must turn in for this assignment.

1. A two paragraph commentary on the work that each team member completed.
2. A description of the eXtensible markup language and its features, strengths, and weaknesses.
3. An explanation of the default serialization method provided by the Java language.
4. The final version of the source code for your serialization benchmarking framework.
5. Output from at least five runs of your benchmarks, demonstrating features and correctness.
6. A detailed report addressing the performance trade-offs associated with serialization.
7. A reflection on the challenges that you faced when completing this laboratory assignment.

In adherence to the Honor Code, students should complete this assignment while exclusively collaborating with the other member of their team. While it is appropriate for students in this class—who are not in the same team—to have high-level conversations about the assignment, it is necessary to distinguish carefully between the team that discusses the principles underlying a problem with another team and the team that produces an assignment that is identical to, or merely a variation on, the work of another team. Deliverables from one team that are nearly identical to the work of another team will be taken as evidence of violating Allegheny College's Honor Code.