Thesis Title

Name

A Thesis in the Field of Information Technology

for the Degree of Master of Liberal Arts in Extension Studies

Harvard University

January 2010

Abstract

The main objective of this project is to ...

Acknowledgements

I would like to thank ...

Contents

List of Figures

List of Tables

List of Equations

List of Algorithms

List of Code

Chapter 1: Introduction

Ever since the discovery of the molecular structure of DeoxyriboNucleic acid (DNA) (Watson & Crick, 1953), the field genetics has undergone a revolution.

...

... the completion of the reference Human genome (International Human Genome Sequencing Consortium, 2001; Venter et al., 2001) and ... mostly single nucleotide polymorphisms (SNPs) ...

...

1.1. Prior Work

...

1.2. Project Goals

The goal of this thesis project is to ...

Chapter 2: Requirements

This chapter specifies the requirements of the system.

...

2.1. High-level Requirements

...

Figure 2.1 depicts the high-level view of data flow through the system.

...

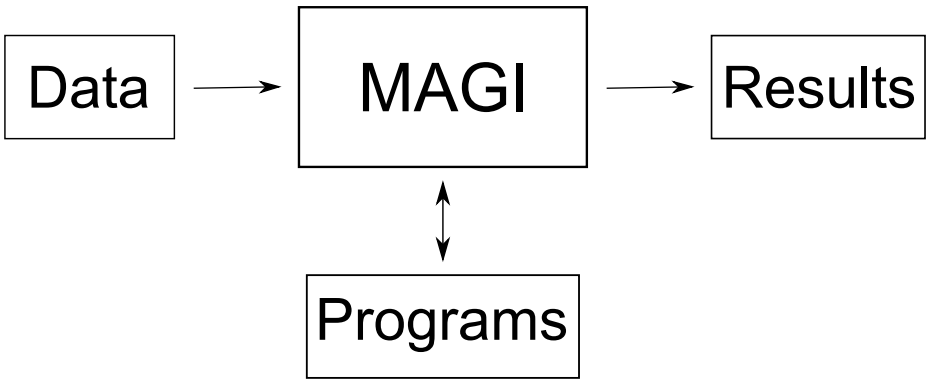

Figure 2.1: High-level view of system functionality

2.2. System Functionality

This section describes in detail the main features and capabilities of the system. The Glossary appendix (Appendix A) describes the important domain concepts.

Chapter 3: Design

3.1. Introduction

...

We defer discussing the implementation details until the Implementation chapter (Chapter 4).

...

Table 3.1 lists the main server technologies we chose for this system.

Type	Technology
Software platform	Java™ Platform, Enterprise Edition
Server implementation	JBoss
Database	Oracle
Object/relational mapping	EclipseLink
Web framework	JBoss Seam

Table 3.1: Server techonologies

Chapter 4: Implementation

We presented the design of the system in the Design chapter (Chapter 3). This chapter describes how we implemented it using Java™ Platform, Enterprise Edition (Java EE). After briefly introducing the key aspects of Java EE in the Java EE Platform section (§4.1), this chapter presents the key implentation details for each tier of the application.

4.1. Java EE Platform

This section introduces the readers to the elements of the Java EE platform to develop and run this system.

4.1.1 Java EE Architecture

...

4.2. Implementation Overview

Figure 4.1 shows the bird's eye view of the objects in this system. There are three distinct tiers: presentation, business, and entity.

...

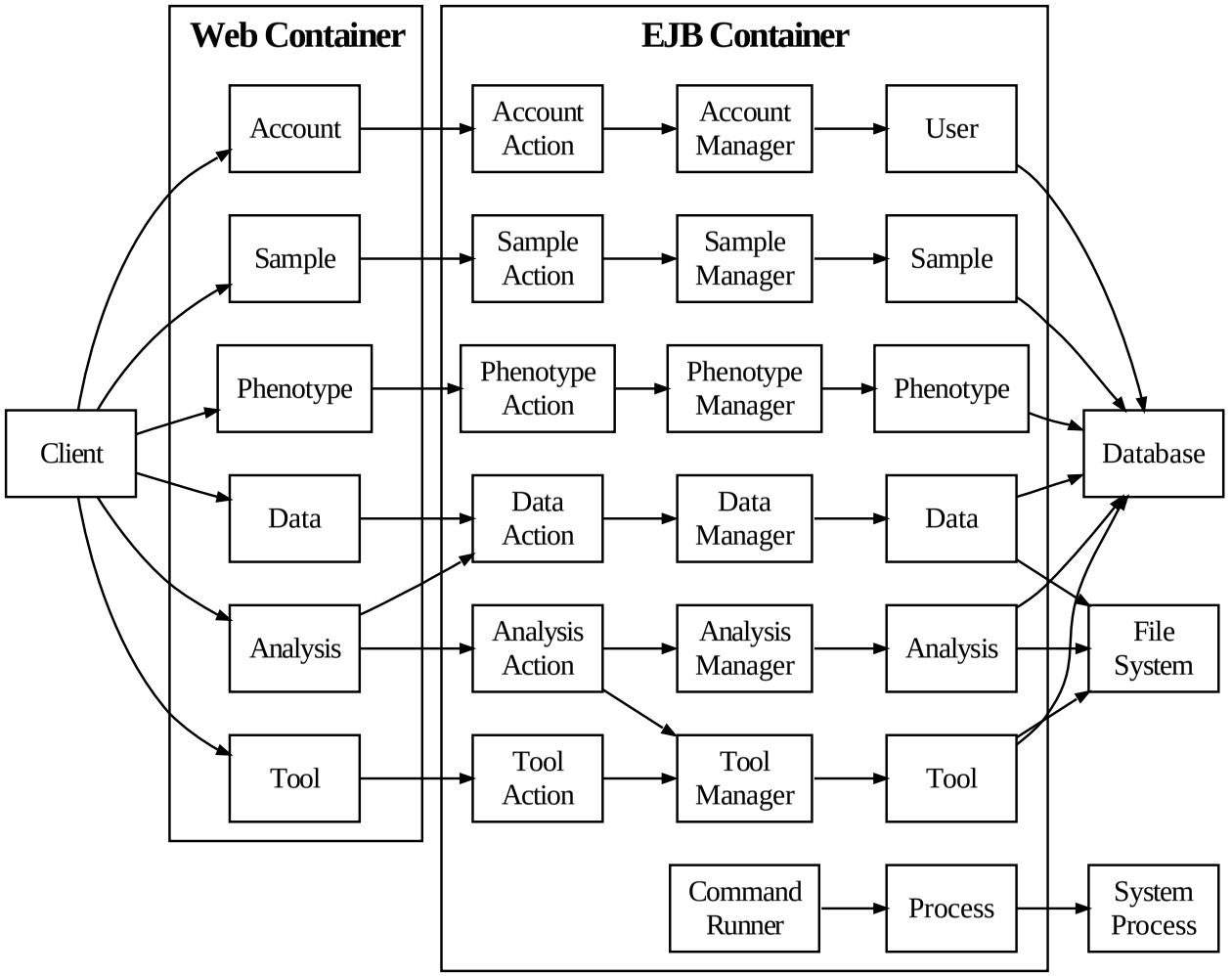

Figure 4.1: High-level object view of the system

...

The code for the Seam component is shown in Listing 4.1.

...

The code for the facelet user interface component is shown in Listing 4.2.

...

```
@Stateful
@Name("sampleSetAction")
@Scope(ScopeType.SESSION)
public class SampleSetActionBean implements SampleSetAction {

  private String sampleSetName;
  private InputStream data;

  @EJB
  private SampleManager sampleManager;
  ...
  public void importSampleSet() { ... }
  ...
}
```

Listing 4.1: An example of Jboss Seam component

```
...
<h:form enctype="multipart/form-data">
  ...
  <h:inputText value="#{sampleSetAction.sampleSetName}"/>
  <s:fileUpload data="#{sampleSetAction.data}"/>
  ...
  <h:commandButton action="#{sampleSetAction.importSampleSet}" value="Submit
    "/>
</h:form>
...
```

Listing 4.2: An example of JSF XTHML component

Chapter 5: Development

This chapter discusses the tools and methodologies employed in the code development of this system.

5.1. Development Tools

For this project we used Subversion (Pilato et al., 2008) as the version control mechanism.

...

5.2. Development Methodologies

...

and Test Driven Development (TDD) (Beck, 2003).

...

Chapter 6: Sequence Analysis Tools and Applications

...

6.1. Needleman-Wunsch Implementation

...

6.1.1 Global Sequence Alignment

...

6.1.2 Longest Common Subsequence (LCS)

...

Essentially the first step of LCS performs this recursion:

$$S_{0,j} = 0 \tag{6.1}$$

$$S_{i,0} = 0 \tag{6.2}$$

$$S_{i,j} = \max \begin{cases} S_{i,j-1} \\ S_{i-1,j} \\ S_{i-1,j-1} + 1 \quad \text{if } A_i = B_j \end{cases} \tag{6.3}$$

...

The pseudocode of LCS is shown in Algorithm 6.1.

...

Algorithm 6.1 LCS($A_{0..n}, B_{0..m}$)

1 $S_{i,j} \leftarrow 0$ **for** all $i = 0$ or $j = 0$ {set first row and first column to 0}
2 $T_{i,j} \leftarrow$ `UP` **for** all i, j {pointing up by default}
3
4 **for** $i \leftarrow 1$ **to** n **do**
5 **for** $j \leftarrow 1$ **to** m **do**
6 $$S_{i,j} = \max \begin{cases} S_{i,j-1} \\ S_{i-1,j} \\ S_{i-1,j-1} + 1 & \text{if } A_i = B_j \end{cases}$$
7 $$T_{i,j} = \begin{cases} \texttt{LEFT} & \text{if } S_{i,j} = S_{i,j-1} \\ \texttt{UP} & \text{if } S_{i,j} = S_{i-1,j} \\ \texttt{DIAGONAL} & \text{if } S_{i,j} = S_{i-1,j-1} + 1 \end{cases}$$
8 BACKTRACE(A, T, n, m)
9
10 **function** BACKTRACE($A_{0..n}, T_{0..n \times 0..m}, i, j$)
11 **if** $i = 0$ or $j = 0$ **then**
12 **return**
13 **if** $T_{i,j} =$ `DIAGONAL` **then**
14 BACKTRACE($A, T_{n,m}, i-1, j-1$)
15 **print** A_i
16 **else if** $T_{i,j} =$ `UP` **then**
17 BACKTRACE($A, T, i-1, j$)
18 **else**
19 BACKTRACE($A, T, i, j-1$)
20 **end**

Chapter 7: Summary and Conclusions

...

In conclusion, I ...

...

7.1. Lessons Learned

There are many lessons learned from the project.

...

7.2. Limitations and Known Issues

...

References

Beck, K. (2003). *Test-driven development : by example*. Boston: Addison-Wesley.

International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(6822), 860–921.

Pilato, C., Collins-Sussman, B., & Fitzpatrick, B. (2008). *Version Control with Subversion*. O'Reilly Media, Inc. `http://subversion.tigris.org/`, retrieved April 2009.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., Gocayne, J. D., Amanatides, P., Ballew, R. M., Huson, D. H., Wortman, J. R., Zhang, Q., Kodira, C. D., Zheng, X. H., Chen, L., Skupski, M., Subramanian, G., Thomas, P. D., Zhang, J., Miklos, G. G. L., Nelson, C., Broder, S., Clark, A. G., Nadeau, J., Mckusick, V. A., Zinder, N., Levine, A. J., Roberts, R. J., Simon, M., & Slayman, C. (2001). The sequence of the human genome. *Science (New York, N.Y.)*, 291(5507), 1304–1351.

Watson, J. D. & Crick, F. H. (1953). Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature*, 171(4356), 737–738.

Appendix A: Glossary

DNA DeoxyriboNucleic Acid, genetic material containing instructions for the functioning of living organisms. They exist in sequences of the four nucleotides `A`, `C`, `G`, and `T`. 1

SNP Single Nucleotide Polymorphism, the most abundant type of DNA polymorphism that involves a single nucleotide and occurs at a particular genomic position. Each SNP is given a name like rs25. The majority of the SNPs have two expected alleles (e.g. rs25 has alleles C and T). 1

Appendix B: Application Code

B.1. Java Code

B.1.1 package account

Listing B.1: /account/business/AccountManager.java

```
/*
 * Proj:  MAGI - A System for Managing and Analyzing Genomic Information
 * Auth:  Huy Nguyen
 */
package com.myapp.magi.account.business;

import javax.ejb.Local;

import com.myapp.magi.exception.DataExistException;
import com.myapp.magi.account.model.User;
import com.myapp.magi.exception.FindSingleException;
import com.myapp.magi.exception.ValidationException;

/**
 * Manager of {@link User}. It provides methods for managing
 * a {@link User User}
 */
@Local
public interface AccountManager {

  public User findByLoginId(String id) throws FindSingleException;

  public void register(User u, Boolean validate)
      throws ValidationException, DataExistException;

  public void remove(User user);
}
```

Listing B.2: /account/business/AccountManagerBean.java

```
/*
 * Proj:  MAGI - A System for Managing and Analyzing Genomic Information
 * Auth:  Huy Nguyen
 */
package com.myapp.magi.account.business;

import java.util.HashMap;
import java.util.Map;

import javax.ejb.EJB;
import javax.ejb.EJBException;
import javax.ejb.Stateless;

import org.slf4j.Logger;
import org.slf4j.LoggerFactory;

import com.myapp.magi.exception.DataExistException;
import com.myapp.magi.account.model.User;
import com.myapp.magi.exception.FindSingleException;
import com.myapp.magi.exception.ValidationException;

/**
 * Manager of {@link User}. It provides methods for managing
 * a {@link User}.
 *
 */
@Stateless(name = "AccountManager")
public class AccountManagerBean implements AccountManager {
  private static final Logger log = LoggerFactory
      .getLogger(AccountManagerBean.class);

  @EJB private UserDAO userDao;

  public AccountManagerBean() {  }

  public User findByLoginId(String id) throws FindSingleException {
    User u = userDao.findByLoginId(id);
    u = userDao.findByLoginId(id);
    return u;
  }

  public void register(User u, Boolean validate)
      throws ValidationException, DataExistException {
    log.debug("register: {}", u);

    if (validate) {
      validateFields(u);

      // check if the user exists by loginId
      try {
```

```
        userDao.findByLoginId(u.getLoginId());
        throw new DataExistException("User exists: " + u.getLoginId());
      } catch (FindSingleException e) {
        // expected
      }
    }
    // TODO: catch SQL error to be informative
    try {
      userDao.persist(u);
    } catch (Exception e) {
      throw new EJBException(e);
    }
  }

  public void remove(User user) {
    userDao.remove(user);
  }

  public void validateFields(User u) throws ValidationException {
    Map<String, String> errors = new HashMap<String, String>();

    if (u.getLoginId() == null || u.getLoginId().length() > 15) {
      errors
          .put("loginId", "Login ID must be between 1 and 15 chars");
    }
    if (u.getFirstName() == null || u.getFirstName().length() > 15) {
      errors.put("firstName",
          "First Name must be between 1 and 15 chars");
    }
    if (u.getLastName() == null || u.getLastName().length() > 15) {
      errors.put("lastName",
          "Last Name must be between 1 and 15 chars");
    }
    if (errors.size() > 0) {
      throw new ValidationException("There are invalid fields",
          errors);
    }
  }
}
```

B.1.2 package utils

...