

OPEN-SOURCE-DATENINTEGRATION FÜR DIE KREBSFORSCHUNG

Pfeifer B¹⁺, Netzer M^{1+*}, Seger M¹, Osl M¹, Baumgartner C¹

Abstract

Ein großer Teil moderner Krebsforschung besteht heute in der Verarbeitung und statistischen Analyse von Daten. Wir verfolgen an der UMIT in aktuellen Krebsforschungsprojekten das Ziel, mit der Verarbeitung und Analyse umfangreicher Daten aus verschiedenen dislozierten Forschungseinrichtungen die Grundlagen für die Entwicklung neuer Therapien und Vorsorge-Maßnahmen zu entdecken. So werden beispielsweise unterschiedliche Cancer-Stage-Gruppen auf neue Behandlungsmethoden überprüft. Ein weiteres Ziel ist, die molekularen Signaturen bestimmter Tumortypen zu identifizieren, um effiziente und schonende Diagnose-Verfahren zu entwickeln.

1. Einleitung

Das Konzept eines Data Warehouses wurde von IBM im Jahre 1980 eingeführt. Zu dieser Zeit wurde ein Data Warehouse als Information Warehouse bezeichnet, und erst im Jahre 1988 wurde der heute gebräuchliche Begriff eingeführt. Wegen der Applikations-orientierten Sicht verwenden viele Hersteller den Begriff Business Intelligence Systeme als Synonym.

Ein Data Warehouse ist eine zentrale Sammlung bzw. ein zentrales Repository zur persistenten Speicherung von analyserelevanten Daten. Versteckt und verdeckt unter enormen Datenmengen liegen wichtige Informationen. Data Warehouse Systemen, welche gekoppelt mit intelligenten Such- und Data Mining Algorithmen sind, erlauben die Transformation der Daten in benutzbare Information und neues Wissen kann generiert werden.

Ein Data Warehouse System „von der Stange“ zu kaufen, wie dies bei Standardapplikationen der Fall ist, ist dabei nicht möglich. Das konzeptuelle Design sowie die Implementierung sind abhängig

¹ UMIT, Institute of Biomedical Engineering, Eduard Wallnöfer Zentrum 1, 6060 Hall/Tirol

* designated speaker, + contributed equally

Berlin Open 09

davon, was man mit dem System machen will, und welche Systeme und Datenquellen im DWH involviert sein müssen [1-4]. Aus diesem Grund kann die Erstellung eines Data Warehouse Systems zu einem kostspieligen Prozess werden. Die Beantwortung von Fragen, welche Daten in das System integriert werden sollen, welchen Impact die Daten auf das Gesamtsystem haben, in welcher Aggregationsstufe die Daten zu integrieren sind, welche Verfügbarkeit externe Datenquellen aufweisen und zu welchen Konditionen diese verwendet werden dürfen, sind von höchster Bedeutung, um ein DWH System erfolgreich aufbauen zu können. Weiters ist zu erwähnen, dass die Anzahl der Datensätze bedingt durch die Historie der Datenbestände in einem DWH System sehr groß sind.

Für gewöhnlich wird bei einem Data Warehouse zwischen den Front- und Backroom Komponenten auf logischer wie auf physikalischer Ebene unterschieden. Während der Backroom die Daten persistent hält und Datenmanagementmechanismen bereitstellt, steht der Frontroom für die Präsentation sowie dem Zugriff auf die Datenbestände bereit. Diese Trennung ist wichtig, um die genaue Funktionsweise eines solchen Systems verstehen zu können. Der Backroom wird desöfteren als Datenmanagement und Datenpreparationskomponente bezeichnet. Hier befinden sich die Daten, werden in das analytische System integriert und zeichnet sich für die Abarbeitung von Anfragen und das Zurückliefern der Daten verantwortlich. Die Frontroom Komponente erlaubt es, dem Benutzer unter Verwendung der BI-Applikationen Anfragen an das System zu stellen.

Eine große Herausforderung bei der Erstellung eines DWH Systems ist, dass der Backroom Spezialist sich mit verschiedenen „Sprachen“ auseinandersetzen muss. Er ist gefordert die verschiedenen Datenquellen zu verstehen, und muss wissen, wie verschiedene Protokolle sowie Betriebssysteme funktionieren. Nur wenn diese Dinge richtig verstanden und umgesetzt wurden, kann ein DWH System als Retter im Datendschungel angesehen werden.

Zusammenfassend lässt sich sagen, dass ein DWH System nicht als eine monolithische Applikation angesehen werden kann, sondern vielmehr ein komplexes interagierendes System darstellt, in dem verschiedene abgestimmte Softwarepakete, Frameworks und Werkzeuge in Konnex gebracht werden.

Eine der meist propagierten Thesen in der Wirtschaft ist: „Zeit ist Geld“. Schnell wachsende Märkte benötigen Werkzeuge, die genaue Analysen ermöglichen, um so wirkungsvoll Unternehmen steuern zu können. In den Bereichen Wissenschaft, Entwicklung, Design, Marketing, Logistik etc. gewinnen Business Intelligence Systeme immer mehr an Bedeutung, und unterstützen die Unternehmensführung die täglichen Herausforderungen zu meistern.

2. Methoden & Ergebnisse

Moderne Techniken wie das elektronische Bearbeiten von Daten erlaubt uns die Nutzung von Informationen wie eine Ressource. Wir befinden uns in einem Informationszeitalter, welches unser Handeln und Tun stark prägt.

Daten selbst können als logisch gruppierte Informationseinheiten betrachtet werden und sind somit fundamentale Einheiten von Information. In der Informatik werden Daten als maschinenlesbare und verarbeitbare digitale Repräsentationen von Information gesehen. Die Information ist dabei Zeichencodiert, die Information wird dabei durch zugrunde liegende Regeln definiert. Daher kann man sagen, dass Information eine Syntax zugrunde liegen hat. Um die Information aus den Daten extrahieren zu können, müssen diese in einem semantischen Kontext interpretiert werden. Als Beispiel könnte man hier die Zahl 43 angeben. Ohne Erklärungen, ohne semantische Einbettung steht die 43 lediglich für eine Zahl, für ein Datum. Wird jedoch ein Kontext gegeben, zum Beispiel, dass 43 die Landeskennung für das österreichische Telefonnetz ist, so wird die Information sichtbar. Diese Umstände sind nicht nur in solch einfachen Beispielen sichtbar, sie sind vielmehr gleichsam in komplexen Datenbeständen vorhanden.

Die heutzutage zur Verfügung stehenden Hochdurchsatzverfahren erlauben das Erfassen von kompletten Genen von Individuen eines Organismus bzw. das Finden von Mustern von tausenden Genen in einer Zelle. Die manuelle Analyse von genomischen, proteomischen und metabolomischen Datensätzen anhand von vorhandener Literatur ist dabei unmöglich. Genau aus diesem Grund ist der Einsatz einer DWH Lösung für Biologen, Biochemiker, Bioinformatikern und Medizinern von eminenter Wichtigkeit.

Molekulare Untersuchungen generieren eine große Anzahl an Daten, welche Gensequenzen, genotyp Korrelationen, Krankheitsverläufe und Krankheitsbilder, oder die räumliche Struktur von Genen abbilden. Die enorme Heterogenität dieser Datenstände wird zusätzlich noch potenziert durch die Bereitstellung verschiedener Datenmodelle und Integrationsprozessen. Diesem Umstand ist es zu verdanken, dass es heutzutage eine große Anzahl an Biodatenbanken gibt. Die geschätzte Anzahl liegt dabei bei rund 1000 [2].

Die in den diversen Forschungsprojekten verwendeten und zu integrierenden Daten sind dabei unterschiedlich aufgebaut. Aus diesem Grund müssen die Daten zuvor klassifiziert werden, damit diese in die Datenintegrationspipeline eingeschleust werden können. Dabei können drei Gruppen

identifiziert werden: strukturierte Daten, semi-strukturierte Daten und vollkommen unstrukturierte Daten [3]. Strukturierte Daten sind dabei am einfachsten zu behandeln, da der Aufbau formal definiert und immerwährend ist. Semistrukturierte Daten sind hingegen nicht vollkommen strikt definiert, ihre Definition ist manchmal unpräzise und unvollkommen. Extensible Markup Language (XML) kann als Beispiel genannt werden, XML ohne eine Document Type Definition (DTD) ist lediglich semistrukturiert. Zu den unstrukturierten Daten gehört die größte Anzahl (ca. 85%) der verfügbaren Daten. In der biomedizinischen Forschung beispielsweise gibt es eine große Anzahl an Publikationen, die genau dieser Klasse angehören. Das automatische Abarbeiten und verstehen von solchen Dokumenten wäre daher von großer Bedeutung in diesem Gebiet. Um solche unstrukturierten Daten in ein DWH System per Extraktion-Transformation-Laden Prozess (ETL) integrieren zu können, bedarf es der Schaffung von geeigneten Extraktions- und Suchroutinen. Können diese Daten zwecks Suche, Analyse oder Validierung nicht integriert werden, so könnte die Information verloren sein, und dadurch kommt es zu einer Verzögerung der Generierung neuen Wissens.

Der automatisierte Vergleich von Millionen von Parametern in Hunderttausenden von Datensätzen soll helfen, Beziehungen zwischen Daten herzustellen. Klar ist: Je mehr Daten man dafür zur Verfügung hat, desto größer ist die statistische Aussagekraft der Ergebnisse. Die von uns betriebenen Forschungsprojekte verfolgen daher das Ziel, seine eigenen Daten in einer Open-Source PostgreSQL Datenbank mit den Daten anderer führender Krebsforschungsinstitute und Kliniken zu verbinden und gleichzeitig auch öffentlich zugängliche Bio-Datenbanken zu nutzen. Das Ergebnis dieser Datenintegration ist ein Data Warehouse für anonymisierte Patientendaten, medizinische Referenzdaten und Genom-Kartografien von enormer Größe und Komplexität.

Die unterschiedlichen Messgeräte und Messmethoden liefern die Daten in heterogenen Formaten. Es handelt sich zum Beispiel um CSV-Dateien, Bilder in hoher Auflösung, RDBMS, XML-Daten, Webservices und selbst generierte Flat Files. Nicht nur die Daten selbst werden integriert, auch die generierten Metadaten, wie die Art der Erhebung und Messung und die Datenquelle.

Für unsere Krebsforschungsprojekte (IMGuS, BIN, HIT) ist es unerlässlich, dass eine Datenintegrationslösung nicht nur mit allen Datenquellen kooperieren kann, sondern auch in der Lage ist, spezielle Verarbeitungsverfahren zu integrieren [8-10]. Dazu wurden eine Reihe von proprietären und quelloffene Lösungen zur Datenintegration evaluiert und es stellte sich heraus, dass das Open-Source-Produkt Talend Open Studio des französischen Herstellers Talend, geeignet

Berlin Open 09

ist [5]. Im Entscheidungsprozess wurden dabei finanzielle, besonders jedoch technische Gesichtspunkte durchleuchtet. Die offene Architektur von Talend Open Studio ermöglicht es universitären Forschungsprojekten, spezielle Komponenten für den Zugang und die Verarbeitung von Daten zu entwickeln.

Ein großer Vorteil des Talend Open Studios ist die perfekt zu bedienende Oberfläche, welche im Hintergrund, je nach Bedarf, Java bzw. Perl Code erstellt. Da unsere gesamte Bioinformatiklandschaft in Java implementiert ist, war dies ebenso mitentscheidend bei der Wahl des Produktes.

Sieht man sich im Bereich Open Source Applikationen, die speziell für den Bereich Life Sciences zugeschnitten sind um, so findest man schnell heraus, dass Java die am meisten benutzte Programmiersprache bzw. verwendete Technologie ist. Rund 29% aller entwickelten Open Source Programme im Bereich Life Sciences sind in Java bzw. zum Teil in Java entwickelt.

Die Erstellung eines Datenzugangs, sowie die Definitionen der Datentransformationen sind eine der zeitraubendsten Tätigkeiten in der klinischen Bioinformatik. Daher musste an diesem Punkt zuerst angesetzt werden. Dabei wurden die Konversionscodes der ETL Prozesse für die Backendsysteme mit dem Talend Open Studio Designer innerhalb weniger Monate in Betrieb genommen. Für die einzelnen Datendomänen (-omics Daten) wurden dabei Lesekomponenten geschaffen sowie die Komponenten für die semantische Zugehörigkeit der Datensätze definiert. Dank des automatisch erstellten Quellcodes konnten danach diese Codes bzgl. eigener notwendiger Modifikationen geändert und in das Integrationsframework eingebaut werden. Ein typisches Szenario für eine Mapping Komponente, die mit dem Talend Open Studio direkt erstellt werden konnte ist in Abbildung1 gegeben.

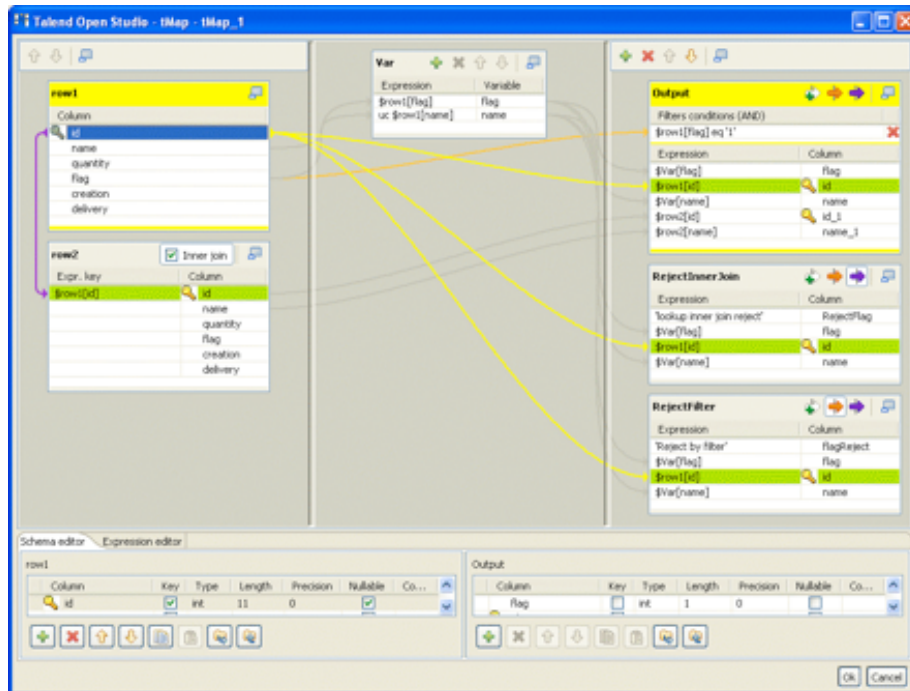


Abbildung 1: Mapping Komponente zur Integration eines Teilbereichs einer Datendomäne in das DWH System

ETL Prozesse zu definieren bedeutet Integrationsworkflows zu erstellen. Auch in diesem Bereich zeigt Talend Open Studio seine Stärken. Die grafische Repräsentation der einzelnen Arbeitsschritte ist perfekt integriert und für den ETL Spezialisten leicht nachzuvollziehen. Durch einen einfachen Mausklick kann zwischen den modularen Integrationsworkflows und dem erstellten Code hin und hergeschaltet werden. Abbildung 2 zeigt ein Beispiel eines ETL Workflows.

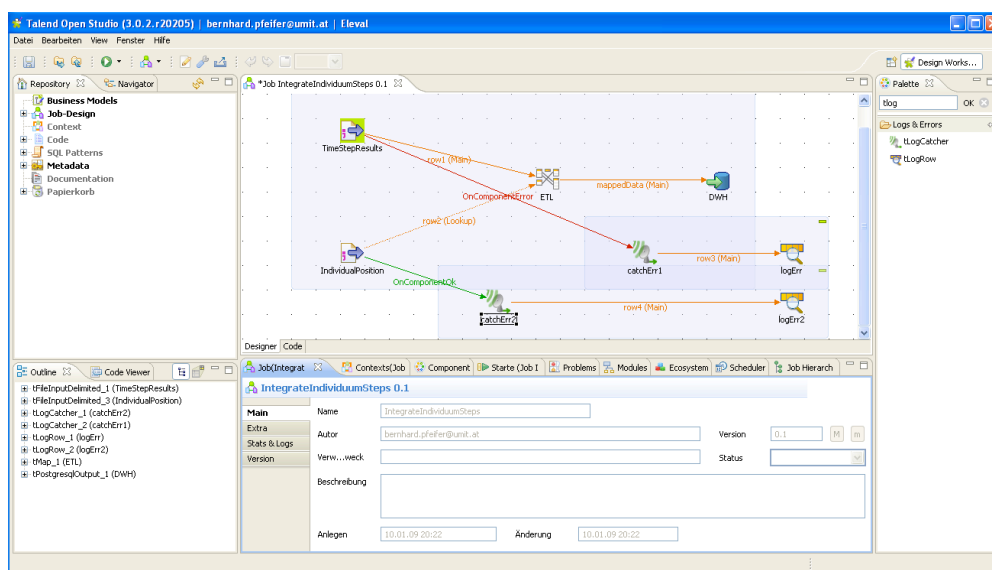


Abbildung 2: Talend Open Studio Designer, Workflow-Perspektive

Das eigentliche DWH trägt dabei den Namen LINDA. LINDA steht für Life Sciences Integrative Data Warehouse, welches auf dem Open-Source-Projekt PostgreSQL basiert. LINDA erhält die Daten in einem zweistufigen Verfahren. Die erste Stufe zentralisiert die Daten aus den vielen unterschiedlichen Quellen. Die zweite Stufe arbeitet die Daten auf, wandelt sie um, bereinigt und ergänzt sie. In dieser Phase werden Referenzdaten aus externen Quellen ergänzt – medizinische Publikationen, Legacy-Systeme, medizinische Referenzdatenbanken. Damit benötigen die Mitarbeiter nur einen Bruchteil der Zeit, die normalerweise notwendig wäre, um Validierungen etc. durchzuführen, da das System speziell auf den Analysebedarf – beispielsweise die Findung von neuen Biomarkerkandidaten – zugeschnitten ist.

Die regelmäßige Aktualisierung des Data Warehouse in jeder Nacht stellt sicher, dass die Forscher ad hoc den neusten Stand der Daten zur Verfügung haben und Data Mining betreiben können. Dies ist aus mehreren Gründen notwendig. Einerseits kommen zum Projekt laufend neue klinische Daten hinzu, und andererseits ändern sich in den externen Datenbanken und Publikationssystemen (pubmed) die Datenbestände täglich.

Darüber hinaus stehen komplexe statistische Verfahren zur Verfügung, um die für ihre Forschungen relevanten Daten zu extrahieren. Auch solche Analysen sind Workflowgetrieben. Hier wird das selbst entwickelte Framework KD3 (Knowledge Discovery in Database Designer) verwendet und bildet den Arbeitsfluss der Forscher genau ab [11]. Es stellt ihnen komplexe statistische Verfahren zur Verfügung und sorgt dafür, dass die Informatiker im Backend und die Krebsforscher im Frontend völlig unabhängig voneinander ihren jeweiligen Aufgaben nachgehen können. Der von Fayyad et al. beschriebene Prozess wird dabei konsequent umgesetzt. Abbildung 3 zeigt dabei den prinzipiellen schematischen Ablauf.

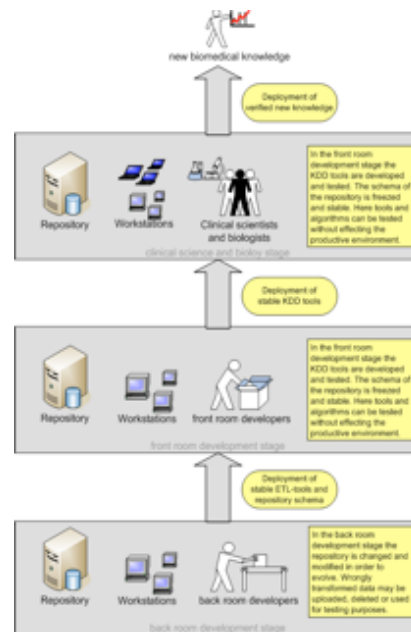


Abbildung 3: Von den Daten des Interesses zu neuem Wissen

Um nicht im Bereich biomedizinischer Statistik das Rad immer wieder aufs Neue erfinden zu müssen, wurde der KD3 hinsichtlich Integration existierender Open-Source Frameworks optimiert. So wurde zum Beispiel das WEKA Statistik Open Source Paket [7] in den KD3 Designer mit Hilfe von Adapterklassen integriert. Abbildung 4 zeigt den KD3 Designer in Aktion.

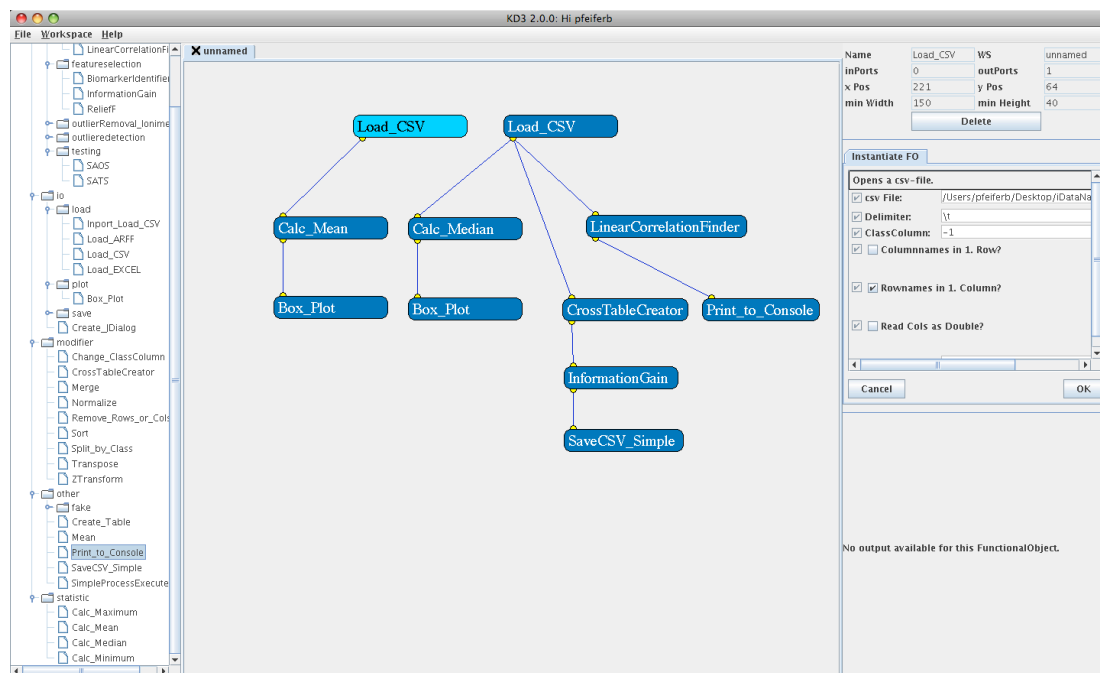


Abbildung 4: KD3 Designer.

Prinzipiell besteht der KD3 auf drei Komponenten. Den „Funktionalen Objekten“, das sind die Elemente, welche zu einem Workflow assembliert werden können, um eine definierte Aufgabe durchführen zu können, dem Workflowdesigner, in dem die Funktionalen Objekte miteinander verbunden werden und einer Parametrierung der Funktionalen Objekte. Der Framework ist in Java implementiert und ist auch in der Lage Codes, welche mittels Talend Open Studio generiert wurden, auszuführen.

Mittels der Verwendung von Talend Open Studio ist es gelungen auf schnelle und unkomplizierte Art und Weise ein System zu schaffen, welches die Bedürfnisse aller Projektteilnehmer zufrieden stellt. Gerade im Bereich der Bioinformatik, wo es notwendig ist, weitere Daten flexibel in das System einzubinden zeigte das Integrationsframework seine ganzen Stärken.

3. Diskussion und Ausblick

Mittels Verwendung von diversen Open Source Produkten ist es gelungen ein professionelles Data Warehouse System aufzubauen, welches den Anforderungen der modernen biomedizinischen Forschung genügt.

Ein wesentlicher Bestandteil des Erfolges ist dabei die Verwendung des Talend Open Source Frameworks, wodurch die Integration aller anfallenden und zur Verfügung stehende Daten auf einfache und schnelle Art und Weise gemacht werden konnte. Die Schaffung der Brücke zwischen den Backroom- (Data Warehouse und ETL Prozesse) und den Frontroom Komponenten (Analyse und Zugriff auf die Daten) abgebildet durch den KD3 Designer, erleichtern es den einzelnen involvierten Forschern sich auf die eigentliche Aufgabe zu konzentrieren: das Finden von potentiellen Biomarkerkandidaten, um die Behandlung und die diagnostischen Verfahren verbessern zu können.

Gerade im universitären Bereich ist die Verwendung von Open Source Produkten unerlässlich und nicht wegzudenken. Das liegt einerseits daran, dass das Geld für Forschungsvorhaben latent knapp ist, und dass geschlossene Systeme einen Eingriff in das System unterbinden und somit die eigentliche Forschung behindern können, da komplizierte Umgehungslösungen gebaut werden müssen.

4. Danksagung

Einen besonderen Dank möchten wir der Firma Talend aussprechen, für die schnelle und kompetente Unterstützung bei Fragen zum Talend-System. Dank auch an die Personen die, die Datenintegration vorangetrieben haben und viele Talend Jobs erstellt, erweitert und integriert haben - Danke an Roland Kienast und Thomas Schwarzmayr. Des Weiteren möchte ich mich bei den Entwicklern des KD3 Designers, bei Herrn Andreas Dander, Michael Handler, Michael Netzer und Leonhard Helminger bedanken, ohne die unsere Analyseplattform nicht möglich gewesen wäre.

5. Referenzen

1. Kimball, R., Caserta, J.: The Data Warehouse ETL Toolkit. Wiley Publishing (2000)
2. Kimball, R., Ross, M.: The Data Warehouse Toolkit. Wiley Publishing (2002)
3. Bauer, A., Gunzel, H.: Data Warehouse Systeme. Dpunkt Verlag (2004)
4. Bloor: Data Warehousing Tools and Solutions. IT-Verlag (1997)
5. Talend: Talend Open Studio (2008)
6. U. Fayyad, G.P.S., Smyth, P.: From data mining to knowledge discovery in databases. Ai Magazine 17 (1996) 37–54
7. WEKA: <http://www.cs.waikato.ac.nz/ml/weka/index.html> (2008)
8. Pfeifer B, Aschaber J, Baumgartner Ch, Dreiseitl S, Modre-Osprian R, Schreier G, Tilg B. A data warehouse for prostate cancer biomarker discovery, BioComp 2007. Las Vegas, USA, 2007. Vol 2: p 316-323
9. Pfeifer B, Aschaber J, Baumgartner C, Dreiseitl S, Modre R, Schreier G, Tilg B. A Life Science Data Warehouse System to enable Systems Biology in Prostate Cancer. 4th International Workshop, p 9ff. DILS 2007, Pennsylvania, USA. 2007.
10. C. Baumgartner, G. Matyas, B. Steinmann, M. Eberle, J. Stein, and D. Baumgartner. A bioinformatics framework for genotype-phenotype correlation in humans with Marfan syndrome caused by FBN1 gene mutations. J. of Biomedical Informatics, 39(2):171-183, 2006.
11. Bernhard Pfeifer, Maria M. Tejada, Karl Kugler, Melanie Osl, Michael Netzer, Michael Seger, Robert Modre-Osprian, Günter Schreier, Bernhard Tilg ; A BIOMEDICAL KNOWLEDGE DISCOVERY IN DATABASES DESIGN TOOL - TURNING DATA INTO INFORMATION; ehealth 2008 Vienna;