UNIVERSITÄTSMEDIZIN GÖTTINGEN : UMG

Pseudonymisierung von Daten in einem Grid

Viezens F, Barz A, Lorberg K

Abteilung Medizinische Informatik, CIOffice Forschungsnetze, Universitätsmedizin Göttingen, Georg-August Universität

Einleitung

Die Einbindung von Komponenten, die eine Pseudonymisierung personenbezogener Daten vor der Verarbeitung in einem Grid vornehmen, ist eine Herausforderung bei der Nutzung verteilter IT-Systeme in der Medizin[1]. Über verfügbare Rechenressourcen kann dem Mediziner eine Grid-Umgebung für rechenintensive Anwendungen zur Verfügung gestellt werden. Der Mediziner soll in die Lage versetzt werden, in einem sicheren Grid- Umfeld genetische und klinische Informationen zu verarbeiten, um zu neuen Erkenntnissen zu kommen. Daher ist ein Workflow [2] der derzeit in anderen Netzen angewendet wird, an eine interne Grid-Umgebung angepasst worden (s. Abb.1). Im weiteren Verlauf soll es für alle Grids gelten.

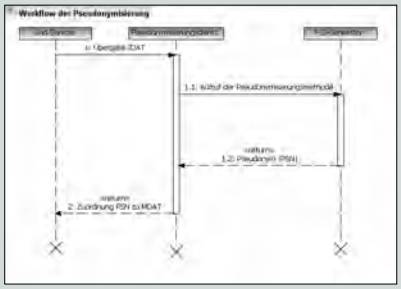

Abb.1: Workflow der Pseudonymisierung im Grid

Materialien und Methoden

In der virtualisierten Teststellung wurde ein Instant-Grid CD-Image verwendet, das innerhalb einer VMWare auf drei virtuellen Maschinen gebootet wurde. Dabei wurde ein Image als Server gestartet und die anderen beiden Systeme als Clients. In der so konfigurierten Testumgebung sollte zuerst ein solch oben beschriebener Pseudonymisierungsdienst selbst geschrieben, im Grid deployed und getestet werden. Im zweiten Schritt ist die Erreichbarkeit eines bereits laufenden Pseudonymisierungsdienstes (Kompetenznetz Angeborene Herzfehler, www.kompetenznetz-ahf.de) von außen zu testen. Das generelle Problem bei einer VMWare besteht in den NAT (Network Adress Translation) - Verbindungen darin, dass die VMWare neben einem „GRID-Master“ ebenfalls einen DHCP-Server für die virtuellen Systeme bereit stellt, der schneller antwortet als der DHCP-Server des GRID-Masters. Die Lösung besteht dabei darin, den GRID-Master mit zwei virtuellen Netzwerkkarten auszustatten. Die erste Netzwerkkarte wird dabei im "Use host-only networking" Modus betrieben. Anschließend werden die Clients gebootet, die im gleichen Modus arbeiten. Erst dann wird die zweite Netzwerkkarte auf dem GRID-Master aktiviert, die ihre Adresse vom DHCP-Server der VMWare bezieht. Über diese Netzwerkkarte erhält der GRID-Master die Möglichkeit fehlende bzw. aktuelle Softwarepakete über das Internet nachzuladen. Die beschriebene Testumgebung ist in der folgenden Skizze (s. Abb.2) dargestellt. Ein weiteres Problem bestand darin, dass die GRID-Server bei all ihren Aktivitäten versuchten, einen über das Instant-Grid fest eingestellten Host mit dem Namen "server" anzusprechen, der in der /etc/hosts nicht korrekt referenziert wurde. Mit einem IP-Alias ließ sich das Problem lösen. Diese Anpassungen waren der Tatsache geschuldet, das das Image eine komplette Filestruktur ausrollte, die für die Funktion der Grid-Beispiele auf der Instant-Grid-CD genügte, aber nicht die kompletten Software-Features beinhalten. Es sind für die GridSphere (www.gridsphere.org)-Portlet (API, JSR 168)-Entwicklung noch Nachinstallationen bei Ant, eine aktuelle Version von GridSphere und Tomcat nötig gewesen [3]. Des Weiteren sind diverse Anpassungen bei der built.xml erfolgt. Der Erfolg beider Ansätze zeigte, dass trotz einigen Aufwandes an Konfigurationen, das auch Legacy-Anwendungen (PSD) in einem Grid-Umfeld integriert werden können. Fehlende Service-Orientierung, sowie der Mangel an standardisierten Schnittstellen konnte erfolgreich ausgeglichen werden. Die Kapselung nach außen erfolgt über eine zusätzliche Schicht.

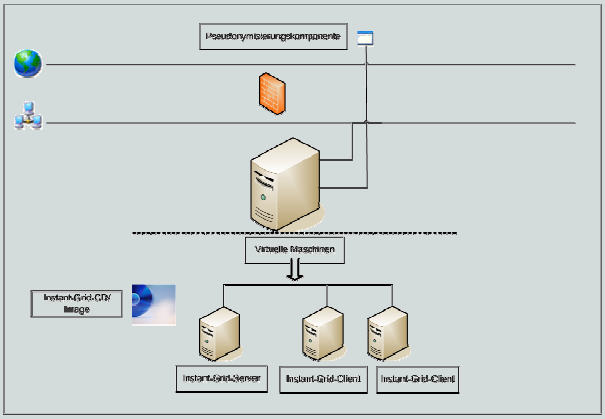

Abb.2: Teststellung der Pseudonymisierung mit einem Instant-Grid-Image in einer Virtuellen Maschine mit getrennten Netzen

Ergebnisse

Ein generelles Problem bei einem Einsatz einer VMWare sind dabei die Netzwerkverbindungen. Die Lösung besteht dabei in der Ausstattung des Servers mit zwei virtuellen Netzwerkkarten. Über die zweite Netzwerkkarte erhält der Master die Möglichkeit Verbindungen nach außen (World Wide Web) zu realisieren, um noch benötigte Software zu downloaden oder Dienste (PSD = Pseudonymisierungsdienst) im Web zu nutzen. Dies wurde hierbei benötigt, um fehlende Komponenten nachzuinstallieren, bzw. die erste Phase der Teststellung zu realisieren. Mit der Einbindung von Pseudonymisierungs - komponenten in einem medizinischen Grid sind die anzuwendenden Datenschutzrichtlinien in Bezug auf die Verarbeitung personenbezogener Daten getestet worden. Dabei wurde in einer geschlossenen Grid-Umgebung, ohne Verbindung nach außen, ein Portlet (s. Abb. 3) generiert, dass diese Aufgabe wahrnimmt. Mit zusätzlich installierter Globus-Grid-Software war dies möglich. Die Filestruktur von Instant-Grid gab diese Möglichkeit des nachträglichen Installierens. Die Funktionsweise wurde erfolgreich getestet (s. Abb. 4). In einer weiteren Phase ist die Erreichbarkeit des PSD vorgenommen worden. Diese Möglichkeit hat weitere Arbeiten an einer Kapselung dieses neuen Features gerechtfertigt. In einem Installationsskript soll diese neue Komponente dem Nutzer bereitstehen.

Abb.3: Erfolgreiches Builden des Pseudonymisierungsportlets

Abb.4: Testergebnis des Pseudonymisierungsportlets

Diskussion und Ausblick

Mit Instant-Grid ist dem Anwender ein Instrument für ad hoc Grid-Anwendungen gegeben. Die Anpassungen mussten im konkreten Fall manuell konfiguriert werden, können aber in Weiterentwicklungen Teil eines Images werden. Das zusätzliche Einbinden von Ressourcen kann auch in einer virtualisierten Umgebung vorgenommen werden. Die zu verarbeitenden Daten (genetische, epidemiologische), durch einen stets erreichbaren Dienst pseudonymisiert, stellen ein Höchstmaß an Datenschutz dar, was stets als Hindernis bei der Verarbeitung medizinischer Daten in konventionellen Grids gesehen wurde[4]. Es werden in allen medizinischen Bereichen zusätzliche Informationen vorhanden sein.

Referenzen

[1] Rienhoff O. Lösungen für sichere Grid-Anwendungen in der medizinischen Forschung. München: Urban&Vogel; 2006:86-90.

[2] Sax U. Stand der generischen Datenschutz-Konzepte sowie deren technischen Realisierung in biomedizinischen Grids. München: Urban&Vogel; 2006:38-43.

[3] Thüne N. Einbindung eines Pseudonymisierungsdienstes in MediGRID. Göttingen: Georg-August-University; 2006.

[4] Viezens F. Grid-Computing in der Biomedizin. München: Urban&Vogel; 2006:56-62.

Diese Arbeit wurde unterstützt durch die D-GRID Projekte MediGRID und Instant-Grid, gefördert durch das Bundesministerium für Bildung und Forschung (BMBF), FKZ 01AK803H/ 01AK807 und dem Kompetenznetz Angeborene Herzfehler (BMBF) FKZ 01G10210.

MediGRID

D-GRID

GEFÖRDERT VOM Bundesministerium für Bildung und Forschung