RC25276 (WAT1203-087) March 29, 2012
Computer Science

IBM Research Report

Accelerating the Deployment of Security Service Infrastructure with Collective Intelligence and Analytics

Maja Vukovic
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 704
Yorktown Heights, NY 10598
USA

Christopher Giblin
IBM Research Division
CH-8803 Rueschlikon
Switzerland

Sriram K. Rajagopal
IBM India
Chennai, India

Research Division
Almaden - Austin - Beijing - Cambridge - Haifa - India - T. J. Watson - Tokyo - Zurich

Accelerating the Deployment of Security Service Infrastructure with Collective Intelligence and Analytics

Maja Vukovic
IBM T.J. Watson Research Centre
Hawthorne, NY 10532, USA
maja@us.ibm.com

Christopher Giblin
IBM Research – Zurich
CH-8803 Rüschlikon, Switzerland
cgi@zurich.ibm.com

Sriram K. Rajagopal
IBM India
Chennai, India
srirraja@in.ibm.com

Abstract—**With the increasing complexity of IT outsourcing environments thousands of servers and their configurations are increasingly managed by globally distributed teams. This requires a flexible identity access management process in place to efficiently provision necessary access rights for a given system, only if users need it, when they need it and for only as long as they need it. In this paper we present a novel approach to discovering required role permissions by integrating system data and enterprise crowdsourcing (a process where a group of experts solve problems through collaboration). By mining server registries, compliance repositories (such as user revalidation records), we derive a set of servers and the respective access rights for each team member. This data is then validated and updated by one or more team members using the principles of crowdsourcing. We show that this approach improves the role discovery process and accelerates the deployment of the security service infrastructure.**

Keywords-component; service delivery, automation, social networking, role engineering

I. INTRODUCTION

With the adoption of the global delivery model, service delivery environments become complex ecosystems. Tens of thousands of system administrators (SAs) are handling thousands of processes that run on thousands of IT systems and their configurations, on behalf of thousands of (outsourcing) customers. To manage theses systems (servers) SAs require access rights. The process for obtaining these access rights, executing them and disposing of them must conform to national and industry compliance regulations.

With the increase number of IT systems managed in delivery centers, that traditional model of individually assigned user IDs is no longer tractable. Traditional paradigm of ID provisioning on endpoints is based on "just in case" model – that someone might need access. This results in a proliferation of user IDs and introduces higher costs and risks. IT service delivery centers and large enterprises nowadays require more effective and flexible privilege user management mechanisms as part of their security and identity management processes.

A novel identity access management system, based on the principles of reusable IDs has been developed to improve the compliance and reduce operational costs for identity access management in large scale data centers [1]. Reusable IDs enable a user to obtain individual user ID for a given system, only if they need it, when they need it and for only as long as they need it. This approach significantly reduces the number of individual admin IDs, and therefore reduces the cost of provisioning and maintaining user IDs, as well as the risk of ID proliferation.

Migrating the Identity and Access Management (IAM) of millions of system administration accounts from a collection of IAM systems to a single role-based model supporting a large, global service delivery infrastructure poses a number of challenges which, given the very large scale, can potentially delay or hinder the infrastructure deployment. Defining roles involves substantial data on systems, security policies, and organization structures. This essential data resides partially within existing inventory and compliance repository silos and partially in human knowledge. The migration also involves the mapping onto a new identity and access management model, in this case Role Based Access Control (RBAC), requiring additional semantics and metadata. Further, global systems and their users are in constant motion.

Collecting both relevant operational data and human knowledge for a global migration must be performed systematically and efficiently with a premium on automation. Coordinating this collection ensures both data and scope, performed in parallel, and decoupled from geography and time zone.

The contribution of this work is the combined use of data analytics and crowdsourcing technology to standardize, automate and consolidate interviews and operational data for the design of access control roles. Subject matter experts (SMEs) are relieved of providing inventories of systems and accounts. Instead, they are asked to verify and refine data, in addition to providing essential information on the characteristics and policies of their local computing environment, information which is not recorded in traditional system inventories and access management systems. Key to this approach is the coordination of the collection of data and knowledge achieved between data integration tooling and an enterprise crowdsourcing platform.

The paper is structured as follows. Next section describes the operational structure of a large service delivery center, the fundamentals of identity access management and role engineering. Section 3 provides the architectural overview of proposed solution detailing on analytics and crowdsourcing components and their interactions. Sections 4 and 5 describe the deployment setup and results respectively. Section 6 puts this work in the context of related efforts in identity access management and enterprise crowdsourcing. Section 7 concludes and identifies future work items.

II. Background

A. Service Delivery Centers

In order to efficiently provide a stack of IT services, a factory model has been applied to the operations in delivery centers. As a result, there are global standardized processes, with adaptive dispatching where activities of services operations are assigned to SAs and SMEs based on the domain and complexity of work. SMEs are grouped into "pools" sharing a specific competency, which is often executed for a set of customers. The knowledge about infrastructure and user practices is captured into common repositories.

B. Identity Access Management

Identity access management system provides a mechanism for the management of individual identities, their authentication, authorization, roles and privileges within or across system and enterprise boundaries with the goal of increasing security and productivity while decreasing cost, downtime, and repetitive tasks. Identity and access management models continue to be challenged by inadequate governance and new delivery models.

C. Role Engineering

Role Based Access Control (RBAC) is a model of access control which was originally formalized in the 1990's and has since been widely adopted and standardized [7]. A role is an entity which associates a set of users with a set of permissions where a permission is the authorization to perform an action on an object. Objects are any information technology resource such as a computer, a file or a printer. A role aligns with a job function and is an indirection allowing user-role and role-permission assignments to be managed separately. The permissions to perform a job function typically change differently than the users who have the job. Thus one of RBAC's key advantages is its simplification of permission assignment compared to managing large numbers of individual user-perrmission assignments. The RBAC model further formalizes the concepts of role hierarchy, session and separation of duty constraint.

While RBAC has proven to simplify the complexity and administration of access control policies, thereby reducing costs and improving security [9], the initial transition to an RBAC system, especially for larger organizations, is recognized as a potentially difficult and costly effort. Defining roles from requires collecting data on complex, often heterogeneous, system environments along with their respective security policies. This information must often be combined or aligned with organizational objectives.

Role mining [10] has been proposed as a means to derive roles from existing access control policies using data mining techniques. Mining operational data alone does not provide a complete role structure. This "bottom-up", data-driven approach must be augmented by a "top-down" design sensibility, incorporating human knowledge not only of the specifics of an IT environment, but of organizational structure and business goals as well [12]. The process of arriving at role structures aligned with business objectives is known as role engineering [8][11].

III. System Architecture

A. Solution Overview

Reusable Ids are credentials in the form of user name and password which are automatically checked-out from an access control server as a system administrator attempts to access a system. Assuming the administrator has rights to the reusable ID, the reusable credential is obtained; the administrator logs on, performs work and logs off, automatically returning the credential to the access control server. This action closely resembles the RBAC model, wherein users activate a role and session. The granularity of permission is that of a system account, namely the reusable ID and its associated group memberships on the target system. The migration of the IAM systems to the global, reusable ID model is therefore a role engineering effort in which common job functions (roles) within the pool are identified and assigned a set of reusable Ids (permissions).

A methodology and technology were developed to support the role engineering effort as part of this migration. Figure 1 provides an overview of the methodology and the technology components involved.

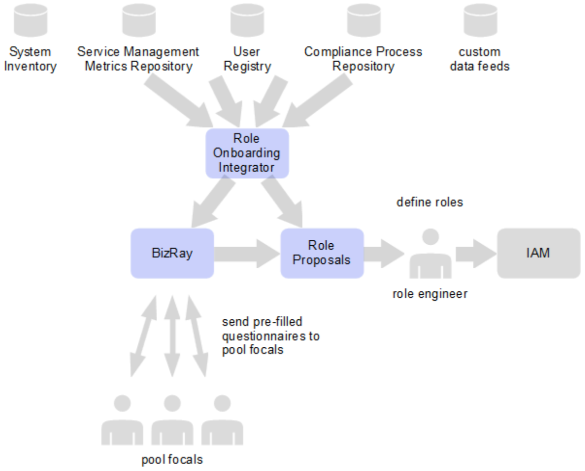

System Inventory
Service Management Metrics Repository
User Registry
Compliance Process Repository
custom data feeds
Role Onboarding Integrator
define roles
BizRay
Role Proposals
IAM
role engineer
send pre-filled questionnaires to pool focals
pool focals

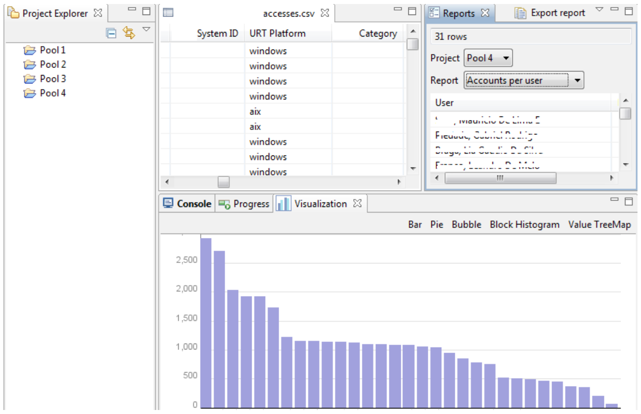
Figure 1: Overview of methodology and architecture

In the initial phase, data pertaining to system accesses is extracted from a variety of repositories, integrated, and cleaned. This data reflects the current state of system accesses and is obtained from the following sources:

- System Inventory (SI) contains information on individual systems such as a unique identifier, hostname, IP address, platform version, and customer account.
- Compliance Process Repository (CPR) stores, in addition to the state of governance processes, a record of all system logins (without passwords) and the system administrators who use them.
- Service Management Metrics Repository (SMMR) maintains data and metrics on pools.
- User Registry (UR) is an LDAP registry of employee information containing a unique identifier, email address and organizational information such as department, title and manager.
- Custom feeds are imported in cases where additional data is required for a specific environment.

Descriptive statistics and visualization support the role engineer in exploring the data, identifying potential inconsistencies. For example, as pools concentrate SME competencies, it is expected that the type and number of system accesses within a pool are evenly distributed. Figure 2 shows the Onboarding Integrator application portraying an uneven distribution of system accesses within a pool, indicating either incomplete data or the need to establish two roles within the pool.

Once data on the current state has been collected, focus shifts to acquiring human knowledge of the target operating environment. This is achieved with a Web-based, enterprise crowd-sourcing platform, BizRay, which automates the distribution, delegation, monitoring and collection of electronic interviews. The interviews elicit information from pool focals on the "to be" state as well as information not yet captured, such as customer account prerequisites. Questions dealing with system and access information are pre-filled with the "as is" data collected during the initial phase. Pool focals therefore do not need to construct an inventory of their pool's accesses, but rather verify and optionally update the pre-filled data; a significant time savings is achieved.



System ID	URT Platform	Category
	windows	
	windows	
	windows	
	windows	
	aix	
	aix	
	windows	
	windows	
	windows	

31 rows

Project Pool 4

Report Accounts per user

User

Figure 2: Distribution of system accesses within a pool as shown by OBI.

After interviews are completed, the refined data on system accesses passes through role-mining technology to propose roles for the pool. The role engineer assesses the role assignments, refining as needed. In the last phase, the finalized role assignments relating users and permissions are exported to the target IAM system for approval.

B. *Questionnaire Structure*

To effectively design roles for each SME in each pool, a set of data about servers, access rights and team members is required. The questions are thematically grouped into:

I. meta data about pool operations (e.g. work activities provided, number and type of customers supported, contact person for reusable ID management, etc.)

II. account list (e.g. password expiration intervals, access type, applications supported, number of IDs, and number of reusable IDs, etc.)

III. account prerequisites (e.g. regulatory properties that must be met by SAs working on these accounts, such as citizenship, drug testing, etc.)

IV. team members and accounts (e.g. team members, their assignment to account and its servers, etc.)

V. data about stepping stone servers (where applicable)

VI. team member details (e.g. email, location, etc.)

C. *Onboarding Integrator*

The Onboarding Integrator (OBI) depicted in Figure 1 is a purpose-built application combining Extract Transform Load (ETL) capabilities with analytics and visualization. Importantly, it implements a REST client for integration with the BizRay crowdsourcing platform. A variety of import and export protocols and formats are supported, most notably, HTTP, LDAP, Comma Separated Value (CSV) files and Excel spreadsheets.

OBI uses a number of technologies for storing and integrating data. Imported data is mapped onto a core model similar to a data warehouse. This model is implemented with the Resource Description Framework (RDF) which allows for flexible extension of data entities, an advantage when linking data from various sources. In some cases where records cannot be joined across repositories with keys, approximate text matching is performed with the Lucene information retrieval library. When performing occasional analysis over millions of tabular records for an entire geography, a relational database is used. These technologies together are arranged architecturally in a data and query layer.

After data for a pool has been imported into OBI, it can be uploaded to the BizRay platform. Data is exported using predefined spreadsheet templates and then uploaded over secure file transfer to the BizRay server. Prior to this transfer, OBI creates a new BizRay interview over the REST interface, assigning the interview to a pool focal entered by the pool assessment manager. BizRay in turns initiates the interview's lifecycle by sending an email to the pool focal, informing of the new interview with the interview's URL. The pre-filled, standardized spreadsheets appear in the interview alongside the appropriate questions. The spreadsheets can be modified by the focal or delegated to other focals, iteratively and collaboratively refining the pool's system permissions.

OBI is implemented as an Eclipse Rich Client (RCP). The RDF triplestore uses the Sesame framework. Text matching is performed with Lucene.

D. *Collaborative Verification*

BizRay [2], shown in Figure 3., is an enterprise crowdsourcing service that accelerates knowledge discovery inside large organizations. Knowledge requests are captured in the distributed questionnaire artifact, consisting of one or more sections, each of which has one or more questions. BizRay manages its lifecycle, similar to a workflow system and facilitates delegation of requests and their subtasking. More than one expert can complete each questionnaire instance. If the data gathered is incomplete or unidentified, the user can forward the request to another expert, requesting their help.

As experts contribute their knowledge, the system keeps track of their identity forming a community around the questionnaire. BizRay triggers reminders and to users who did not respond to the request in a given time.

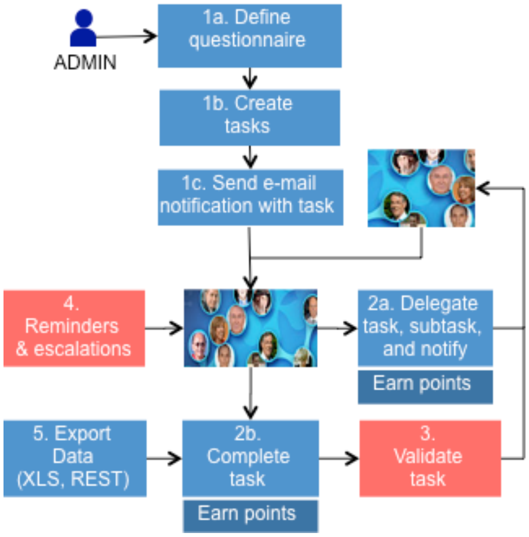

ADMIN
1a. Define questionnaire
1b. Create tasks
1c. Send e-mail notification with task
4. Reminders & escalations
2a. Delegate task, subtask, and notify
Earn points
5. Export Data (XLS, REST)
2b. Complete task
Earn points
3. Validate task

Figure 3. Main BizRay capabilities

We now turn into internal representation and structure of BizRay questionnaires, and discuss its interaction with analytics module. BizRay defines a generic “SurveyResponse” data object, shown in Figure 4, to save responses across various questionnaires. Each questionnaire aims to collect information organized around a particular subject and tasks are created against the primary users’ names who may have such information. For example, if an enterprise plans to collect information about tools used in various projects, a task shown in Figure 5, is created against each Project Manager to capture information about their project.

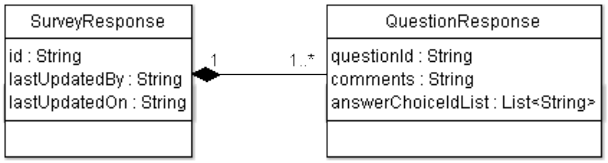


Figure 4. Artifacts capturing questionnaire structure

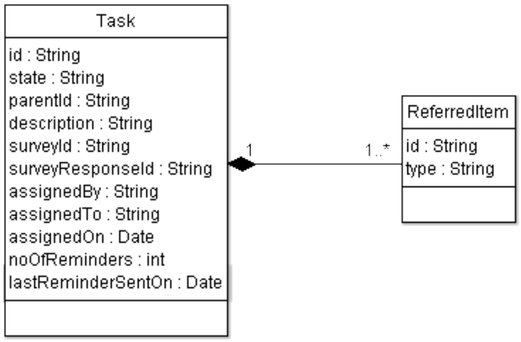


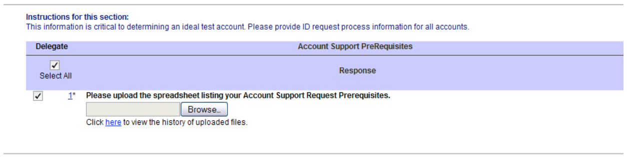
Figure 5. Artifacts capturing task properties

Each parent task is associated with an instance of the "SurveyResponse" data object identified by its ID. Each task contains a reference to the questionnaire – surveyId and the response – surveyResponseId. In addition, it contains the details such as who assigned the task (assignedBy) to whom (assignedTo) and when (assignedOn). The details of the questions referred to the assignee in the task are stored in the supporting "ReferredItem" object. BizRay allows multiple users to collaborate on the same task by supporting creation of sub tasks. The parent task and all its sub tasks share the same instance of "SurveyResponse". The parent task is identified by the parentId attribute of the "Task" object.

Responses to questions of different types (such as text, text area, single/multiple choice, etc.) are saved in the supporting "QuestionResponse" data object. Files uploaded by users against "Upload" type questions are saved on the file system. They use the following naming convention: <Survey Response Id>+<Question Id>+<User Id>+<Time stamp>+<File extension> and can be found in the folder named <SurveyID>, after the ID of the questionnaire. Such a naming convention helps BizRay quickly locate the files uploaded by users for a particular question for a particular response. Users can upload multiple versions of the file and to maintain a history of the same, the user ID and the timestamp are captured. During report generation, only the most recent uploaded file is considered.

BizRay exposes REST-based Web services, which can be used by authenticated clients to create tasks against any questionnaire. Tasks are created for unique entity instances. For example, if tasks are created to gather information about platforms used in a project, the entity instances would be "Windows", "Unix" etc. When the client tries to create a new task for an existing entity instance, a new sub task is created under the parent task corresponding to the entity instance. The Web service also allows the client to pre-populate responses for a task. This is useful in scenarios where part of the information to be collected is available in data sources within the organization. This enables users to simply validate information available in most cases. To pre-populate "Upload" type questions, BizRay server allows secure FTP access to authenticated clients. The client can create the desired file using the naming convention discussed above and upload to an appropriate folder from where it is picked up when a user opens the task.

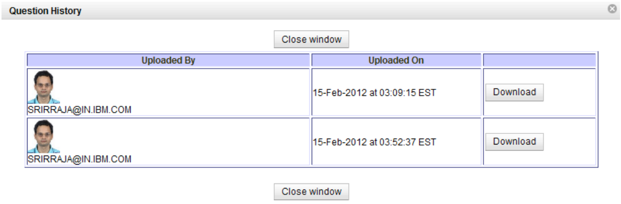
Figure 6 shows a sample "Upload" type of question, while Figure 7 demonstrates how user can view the history of uploaded files and download desired copies for a given question.


Instructions for this section:
This information is critical to determining an ideal test account. Please provide ID request process information for all accounts.

Delegate	Account Support PreRequisites
☑ Select All	Response
☑ 1*	Please upload the spreadsheet listing your Account Support Request Prerequisites.

Click here to view the history of uploaded files.

Figure 6. Upload type of question


Question History

Uploaded By	Uploaded On	
SRIRRAJA@IN.IBM.COM	15-Feb-2012 at 03:09:15 EST	Download
SRIRRAJA@IN.IBM.COM	15-Feb-2012 at 03:52:37 EST	Download

Figure 7. History of uploaded files to a single question

IV. ROLE DISCOVERY PROCESS AND DEPLOYMENT

A. *Deployment Setup*

Given the sheer number of customer accounts and servers managed on their behalf in the service delivery centers, for deployment purposes they are grouped on geographic basis. Depending on the cost and compliance requirements delivery centers may manage and support customers even from multiple geographies.

For the purpose of IAM deployment, the pool assessment was performed on the Geo-basis. Selected delivery center in three regions: Americas (AG), Europe, the Middle East and Africa (EMEA) and Asia Pacific (AP) were targeted. Each Geo had a deployment lead assigned to it who work together with a set of pool assessment managers (PAMs) to reach out to pool focals (PFs) and coordinate the pool assessment and IAM deployment process. Pools provide system administration functions, for different platforms (e.g. Unix, Intel), for a set of customers in a single Geo. There are specific cases where there are pools supporting customers in multiple Geos, which adds complexity to provisioning and storing access rights due to variations in privacy regulations.

The first wave of role discovery was executed predominantly in a manual manner. Geo deployment leads would group pools and assign them to PAMs, who would then e-mail required spreadsheet templates to each PF. PFs would coordinate data gathering process within the pool. In case PF was no longer overseeing a specific pool, PAM had to track that and track the chain of e-mails and follow-up individually with each new PF or their delegate.

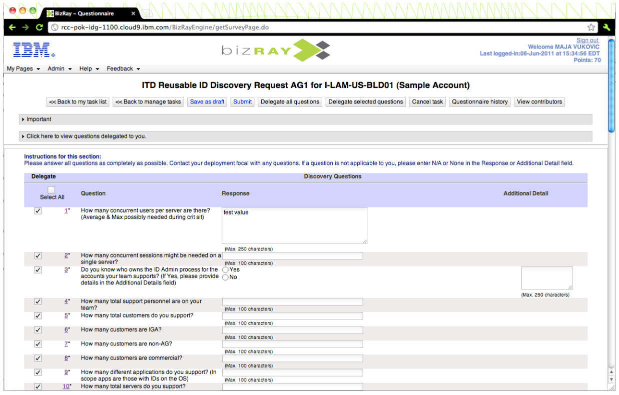
Despite of, and in addition to, existence of centralized repositories of servers, access rights and assigned pool members, pools would often track their server ownership in a custom manner (e.g. CSV, XLS files, or using alternate server repository services). Often when communicating back the required data to PAMS, they would supply non-standardized server listings. This would further increase the complexity of task for PAMs, who would have to resolve the data inconsistencies, manually track the responses and coordinate PFs.

B. *Deployment Results*

To accelerate the discovery of server and access right data, we have developed an analytics module to gather the required data from enterprise data sources. The data is formatted based on the pool assessment requirements, and sent to PFs to validate it, thereby simplifying both their and PAMs tasks. Table 1 summarized the results of pool assessment using crowdsourcing only, and using crowdsourcing with analytics.

1) *Deployment results without Onboarding Integrator*

In the first iteration we have deployed only the questionnaires to the pool focals, without enabling the onboarding integration tool.



ITD Reusable ID Discovery Request AG1 for I-LAM-US-BLD01 (Sample Account)

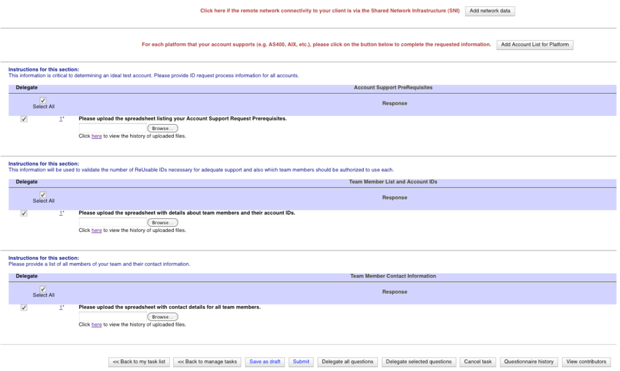
Instructions for this section:
Please answer all questions as completely as possible. Contact your deployment focal with any questions. If a question is not applicable to you, please enter N/A or None in the Response or Additional Detail field.

Delegate □ Select All		Discovery Questions Question	Response	Additional Detail
☑	1*	How many concurrent users per server are there? (Average & Max possibly needed during crit sit)	test value (Max. 250 characters)	
☑	2*	How many concurrent sessions might be needed on a single server?	(Max. 100 characters)	
☑	3*	Do you know who owns the ID Admin process for the accounts your team supports? (If Yes, please provide details in the Additional Details field)	○ Yes ○ No	(Max. 250 characters)
☑	4*	How many total support personnel are on your team?	(Max. 100 characters)	
☑	5*	How many total customers do you support?	(Max. 100 characters)	
☑	6*	How many customers are IGA?	(Max. 100 characters)	
☑	7*	How many customers are non-AG?	(Max. 100 characters)	
☑	8*	How many customers are commercial?	(Max. 100 characters)	
☑	9*	How many different applications do you support? (In scope apps are those with IDs on the OS)	(Max. 100 characters)	
☑	10*	How many total servers do you support?		

Figure 8. Pool assessment questionnaire without OBI integration

Figure 8 shows the sample questionnaire, consisting of 10 section. First one was Web-based requiring for input of pool meta-data, the others required user to upload the XLS-based files, with server listings and access rights. Before pool focals started using the tools, pool assessment managers were provided with training. PAMs then in turn have briefed PFs. Deployment of BizRay for first set of pool assessments, automation of spreadsheet dispatch and coordination of responses, has introduced labor savings. 32 pools in a single region used to tool to provide the required data. The job of PAMs has been simplified, as the tool was automatically dispatching the tasks and their reminders. In addition, full audit trails have been proven useful to trace multiple file uploads and changes to data. Deployment leads now had a central, point-in-time, overview of the pool assessment process, and no longer had to manually reach out to PAMs. The main challenge was collecting the server repository listings from PFs, who would typically have it available in various formats.

2) *Deployment results with Onboarding Integrator*



Click here if the remote network connectivity to your client is via the Shared Network Infrastructure (SNI)

For each platform that your account supports (e.g. AS400, AIX, etc.), please click on the button below to complete the requested information.

Instructions for this section:
This information is critical to determining an ideal test account. Please provide ID request process information for all accounts.

Delegate ☑ Select All		Account Support PreRequisites Response
☑	1*	Please upload the spreadsheet listing your Account Support Request Prerequisites. Click here to view the history of uploaded files.

Instructions for this section:
This information will be used to validate the number of ReUsable IDs necessary for adequate support and also which team members should be authorized to use each.

Delegate ☑ Select All		Team Member List and Account IDs Response
☑	1*	Please upload the spreadsheet with details about team members and their account IDs. Click here to view the history of uploaded files.

Instructions for this section:
Please provide a list of all members of your team and their contact information.

Delegate ☑ Select All		Team Member Contact Information Response
☑	1*	Please upload the spreadsheet with contact details for all team members. Click here to view the history of uploaded files.

Figure 9. Pool Assessment Questionnaire – With OBI

In the second iteration we have integrated OBI to automatically prepopulate the questionnaires, shown in Figure 9, so that the PFs only need to complete the first Web-form section and review uploaded data. The benefits were three-fold: reduction in the human effort, human error prevention and consistent report formats. This setup was deployed n 5 delivery centers in EMEA region.

Table 1 shows deployment results, demonstrating that the average completion time has been significantly reduced once OBI has been introduced in the process. Within only 7 weeks of running in the OBI-enabled mode 75% of tasks are partially completed. Also the average task completion time has been significantly reduced by 65% to date. Note: all task counts include both parent and child tasks (subtasks, referred to as delegated tasks).

Metric	**Without OBI**	**With OBI**
Average completion*	35 days	12.3 days
Num. of questions	20	23
Num. of created tasks	32	31 (70 planned)
Num. of completed tasks*	43	18
Num. of delegated tasks	9	4
Num. of partially completed tasks	0	16
Num. of reminders sent	1	0
Deployment duration	6 months 06/11-02/12	7 weeks 02-03/2012

Table 1. Response results with and without OBI

Average completion time is computed by comparing the timestamps between the first task access and its submission. It many cases, users didn't spend actually 35 days working on the task, but rather may have logged in, answered the questions that they could, and came back after receiving a reminder to complete the task.

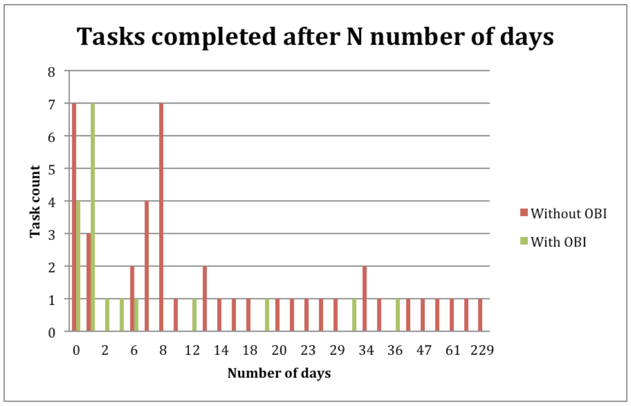


Figure 10. Task and subtask duration to completion

Another important consideration is that deployment was not always linearly progressing. Especially in the case without OBI, the leads may have put some of the discovery efforts on hold temporarily to focus on activities with higher priority, and may not have proactively sought responses from PFs.

Figure 10 shows the distribution of responses in both deployments setups. In the group that used the system with OBI we can observe that within the first business week 20% more tasks were completed.

V. Discussion

Whilst crowdsourcing has been demonstrated as an effective approach to knowledge discovery within the enterprise [4], it does not guarantee significant productivity improvement as several “human factors” issues still persist. Firstly, there is a slow transition and unpredictability element in such campaigns. Due to time-constraints, experts need to quickly grasp the business importance, value and cost of their knowledge contribution. Secondly, there is a throughput limitation in the top-down approach of these campaigns. Thirdly, there are knowledge quality and aging issues. Experts may provide a fast, rather than correct response, or may not be aware of the ground truth at all. Furthermore knowledge acquired today may no longer be valid later during the deployments (e.g. servers may be decommissioned). Incentives make all the difference between success and failure of crowdsourcing campaigns. Over the time the challenge becomes to retain the expert contributors.

As is recognized for any large analytics undertaking, integrating multiple data sources is powerful yet arduous. The ability to link records between different databases while dealing with noise due to missing or inconsistent field values are common problems. Data often resides within repository interfaces and security policies, which serve the repository's purpose very well but complicate data extraction and impact download performance for analytics applications. Solutions to these problems are usually found but may be hard-won.

For the software development a responsive, agile process was adopted in order to work closely and iteratively with pool assessment managers and pool focals. This has been indispensable in dealing with unexpected changes and adapting to local peculiarities. A community of users has been fostered through the use of an online discussion forum. Web conferences have been used for training of the diverse user population.

VI. Related Work

Computing applications are more so than ever relying on human intelligence and skills to execute computational tasks. This method, often referred to as crowdsourcing, is a Web-based approach for employing the collective intelligence of a large network of individuals.

Enterprises, primarily driven by the promise of a low-cost and scalable workforce, are harvesting online experts for a variety of atomic tasks, such as data transcription and classification, to more complex ones, such as generating new business ideas and solving research problems [3]. More recently enterprises turned into their own in-house crowds to improve internal business processes. Crowdsourcing has made scalable reach to large networks of SMEs possible.

Vukovic et al. [4] demonstrated the value crowdsourcing campaigns within the enterprise firewall in terms of process improvement, labor avoidance and discovery of expert communities. Using crowdsourcing mechanism in-house SMEs were engaged to capture the infrastructure information required as part of IT optimization activities.

Stewart et al. [5] used crowdsourcing to effectively tap into the collective intelligence of multilingual employees to translate sentences or correct machine translated sentences for improving translation accuracy and quality of results provided by an automated translation system.

Similarly, Smith [6] deployed a productivity game to engage employees in a large global enterprise, and use their extra time and language abilities to assess localized versions of the Windows operating system in different languages.

To realize the potential benefits of enterprise crowdsourcing, challenges remain. Firstly, how can we effectively engage SMEs and provide suitable incentives. For pool assessments it was assumed that PFs and SMEs would provide the data, however, as they deal with high priority client support requests at the same time, we need right motivation mechanism in place. The first step is communicating the larger business objective (i.e. compliance improvements) that is driving the knowledge discovery effort, and well as the level on the micro – task level. The second step is communicating the actual time effort and designing the rewards to motivate contributions of first-timers and subsequent participants.

Integrating top-down and bottom-up approaches is a common theme throughout the role engineering literature [11][12]. This is the first work, which demonstrates the automated bridging of both approaches by integrating role analytics tooling with a crowdsourcing platform.

VII. Conclusions and Future Work

We presented a novel approach to discovering role permissions in large service hosting environments based on the principles of analytics and collective intelligence. Our system mines server and compliance repositories to derive existing access rights for system administrators. Once SMEs collectively verify this data, our system computes and onboards the discovered roles as part of the rollout of new security service infrastructure. We deployed our approach to 63 pools in two different regions. Our results show that the integrated approach introduces 65% process improvement in the knowledge discovery that is the essential part of on boarding procedure, compared to the practice that relied on SMEs to both collect and verify the data.

Developing a comprehensive framework to support steady state operation of the security service infrastructure remains as a key goal for the future work for the role-mining component. As we improve the quality of the data sources and their value, we will develop mechanisms to automatically regulate health of the roles.

Richness and complexity of an individual human processing element make it challenging to predict the emergent behavior arising from the interactions of a large network of experts in collective intelligence applications in the enterprise domain. This further deepens the problem of reaching desired global behavior (e.g. consistent and timely responses across all the pools) by designing incentive schemes for coordinated problem solving among individual expert (e.g. pool focals and SMEs) who follow their own goals.

Acknowledgments

The work described in this paper was conceived and developed together with Marcel Graf. We acknowledge and thank Craig Ferguson, Milton Hernandez and Paolo Scotton for their valuable contributions. We also acknowledge and thank IBM Technology Delivery for the excellent collaboration and the support of this project.

References

[1] K. Bhaskaran, M. Hernandez, J. Laredo, L. Luan, Y. Ruan, M. Vukovic, P. Driscoll, D. Miller, A. Skinner, G. Verma, P. Vivekanadan, L.Chen, G. Gaskill, "Privileged Identity Management in Enterprise Service-Hosting Environments**"**, Proceedings of Network Operations and Management Symposium **(**NOMS 2012)

[2] J. Laredo, M. Vukovic, S. Rajagopal, "Service for Crowd-Driven Gathering of Non-Discoverable Knowledge". Proceedings of International Conference on Service Oriented Computing (ICSOC '11). December 2011.

[3] D. Brabham, Crowdsourcing as a model for problem solving: An introduction and cases. Convergence, The International Journal of Research into New Media Technologies, Volume 14(1), pp75–90. 2008.

[4] M. Vukovic, M. Lopez, and J. Laredo. People cloud for globally integrated enterprise. In International Conference on Service Oriented Computing.1st International Workshop on SOA, Globalization, People, and Work. 2009.

[5] O. Stewart, J.M. Huerta, M. Sader, Designing crowdsourcing community for the enterprise. In Proceedings of the ACM SIGKDD Workshop on Human Computation (HCOMP '09).

[6] R. Smith, Using games to improve productivity in software engineering. In Proceedings of the ACM SIGKDD Workshop on Human Computation (HCOMP '10)

[7] ANSI. American national standard for information technology – role based access control. ANSI INCITS 359-2004, February 2004.

[8] E. J. Coyne. Role engineering. In RBAC '95, ACM, 1996.

[9] NIST, National Institute of Standards and Technology, 2010 Economic Analysis of Role-Based Access Control, December 2010.

[10] M. Kuhlmann, D. Shohat, and G. Schimpf. Role mining - revealing business roles for security administration using data mining technology. In SACMAT '03, ACM, 2003.

[11] E.J. Coyne, J. M Davis, Role Engineering for Enterprise Security Management. Artech House Publishers. 2007.

[12] S. Calo, C. Giblin, M. Graf, G. Karoth, J. Lobo, I. Molloy, A. Wespi, Towards an Integrated Approach to Role Engineering. In SafeConfig '10 Proceedings of the 3rd ACM workshop on Assurable and usable security configuration, 2010.