Towards Cloud-based Analytics-as-a-Service (CLAaaS) for Big Data Analytics in the Cloud

Farhana Zulkernine[1], Patrick Martin[1] and Ying Zou[2]
[1]School of Computing,
[2]Dept. of Electrical and Computer Engineering
Queen's University
Kingston, ON, Canada K7L 3N6
[1]{farhana, martin}@cs.queensu.ca,
[2]ying.zou@queensu.ca

Michael Bauer[1], Femida Gwadry-Sridhar[2]
[1]Dept. of Computer Science,
[2]Dept. of Medicine Pharmacology and Physiology
Western University
London, ON, Canada
[1]bauer@csd.uwo.ca, [2]femida.gwadry-sridhar@lhsc.on.ca

Ashraf Aboulnaga
Dept. of Computer Science
University of Waterloo
Waterloo, ON, Canada N6A 3K7
ashraf@cs.uwaterloo.ca

***Abstract*—Data Analytics has proven its importance in knowledge discovery and decision support in different data and application domains. Big data analytics poses a serious challenge in terms of the necessary hardware and software resources. The cloud technology today offers a promising solution to this challenge by enabling ubiquitous and scalable provisioning of the computing resources. However, there are further challenges that remain to be addressed such as the availability of the required analytic software for various application domains, estimation and subscription of necessary resources for the analytic job or workflow, management of data in the cloud, and design, verification and execution of analytic workflows. We present a taxonomy for analytic workflow systems to highlight the important features in existing systems. Based on the taxonomy and a study of the existing analytic software and systems, we propose the conceptual architecture of CLoud-based Analytics-as-a-Service (CLAaaS), a big data analytics service provisioning platform, in the cloud. We outline the features that are important for CLAaaS as a service provisioning system such as user and domain specific customization and assistance, collaboration, modular architecture for scalable deployment and Service Level Agreement.**

Keywords- Analytics, workflow, taxonomy, service, CLAaaS, AaaS, cloud, scientific workflow management system, analysis

I. INTRODUCTION

Data analytics has proven its potential in providing decision support in financial, administrative, and scientific sectors by enabling complex computations to generate knowledge, insights, and experimental proofs for scientific discovery. However, the amount of data that needs to be analyzed is growing at an exponential rate and the experts on data analytics technology such as data mining and machine learning often do not have the required domain knowledge to understand the data that needs to be analyzed. Therefore, researchers have been working on designing analytic systems to facilitate complex data analysis. However, these systems are often catered for specific data domains and do not provide ubiquitous access or the desired scalability for big data analysis. The cloud computing technology offers solutions to the above shortcomings but a service infrastructure must be in place to provision the necessary resources transparently on the cloud [13].

The term *"Analytics"* represents a broader scope that includes analysis and inference techniques for decision support. Some of the analytic jobs are *definitive* and are geared towards specific outcomes which are fed into decision support systems, while others are more *exploratory* where the end results may or may not be useful. The exploratory analytic jobs are dynamic in nature and typically require several revisions before they can be used as definitive jobs [11]. A data analysis process [14][17], which we call an *Analytic Workflow,* generally comprises a sequence of data cleansing and integration tasks and/or exploratory analytic jobs such as definition and execution of machine learning models or simple analytic queries. A task in a workflow can also be another workflow for more complex multi-level analytics. A workflow management system is, therefore, mandatory for efficient definition and execution of analytic workflows. The software tools for the tasks and analytic jobs and the visualization models vary depending on the data type, size, domain, and business goals. The proximity of the analytic tools to the data sources is important especially for big data, to avoid data transfer time and network cost. Selecting the appropriate hardware and software resources and defining data and task control flows with dependencies can be a challenge even for an expert data scientist [13]. Therefore, the cloud-based infrastructure should not only provide scalable hardware resources but also a platform equipped with customizable domain-specific software tools and a workflow management system to facilitate the definition and execution of big data analytic workflows [8][9].

A number of analytic workflow management systems exist today, which are also referred to as scientific workflow management systems in the literature [4][16][22][26]. Many of these systems evolved from domain specific analytic

978-0-7695-5006-0/13 $26.00 © 2013 IEEE
DOI 10.1109/BigData.Congress.2013.18

research studies. The systems vary in their focus on data domain, workflow types, representations, execution and provenance mechanisms, and support for collaboration and visualization. Taverna and XBaya [5] support composition and distributed execution of workflows using 3rd party web and data services. With respect to the analysis of large data sets, grid technology is used for distributed mapping and enactment of workflows by WINGS, VLAM-G, DAGMan and Kepler [5]. However, grids are not as scalable as clouds. Cloud can provide ubiquitous access, on demand storage, memory and compute resources, fault tolerance and online collaboration. Distributed and parallel computation frameworks such as Hadoop [10], and the scalable big data storage, management and query tools [9] make the cloud an excellent platform for analytics. Although data security is a major concern for the cloud paradigm [3], techniques such as access control [6], intrusion detection [1], data anonymization [6] and encryption [27] are being used and researched as possible solutions.

None of the existing analytic software and workflow systems [4][5][16][26] meet the desired accessibility and scalability, and provide the customizability to support various user roles such as analysts, domain experts, workflow executors or simple query executors. Considering the required features of an analytic workflow system as discussed above, we propose the conceptual architecture of CLAaaS, a Cloud-based Analytics-as-a-Service (AaaS) platform for big data analytics. As *Platform-as-a-Service (PaaS)*, CLAaaS, will provide on demand data storage and analytics services through customized user interfaces which will include query, decision management, and workflow design and execution services for different user groups. CLAaaS will apply Service Level Agreements (SLAs) to provide controlled access to domain specific software and data resources, and recommendations and guidance in designing, sharing and executing analytic workflows. We use a taxonomy based on a study of existing workflow systems to identify the key features to be included in CLAaaS.

The rest of the paper is organized as follows. Section 2 includes a study of the related work. A taxonomy of analytic workflow systems is presented in Section 3, which is used as to identify key features and requirements for designing CLAaaS. Section 4 describes the conceptual architecture of CLAaaS. The paper concludes in Section 5.

II. Related Work

We propose CLAaaS as a service provisioning platform or PaaS and not as a single analytic software or system. CLAaaS will be configured with one or more analytic software, and most importantly, an analytic workflow management system based on the SLA. There are a number of different analytic software and workflow systems [4][5][16][22][26], which we discuss in this section as related work. SAS [21], SPSS, Cognos, InfoSphere BigInsights [12], and Tableau [23] are a few commercial products for statistical, business and scientific data analysis. Most of them provide rich tools for text analysis, analytic modeling, predictive analytics, visualization, collaboration, decision management, or adding 3rd party applications for decision support.

R, which is a successor of S, is a popular open source software suite used to develop programs to perform statistical analysis [19]. Several open source data mining tools such as Weka [25] and RapidMiner [20] are used widely in various research domains. RapidMiner includes some libraries of Weka and provides graphical user interfaces (GUI) for designing simple processes. However, all the above systems currently need to be installed on organizational frameworks and are not offered as services.

Google Trends and Analytics [9] is an online analytical service, which allows mainly tracking of search keywords from user inputs for various ecommerce applications. The statistics can be used to understand consumer demands and advertise products better. Google Trends also supports graphical visualization.

Several analytic workflow systems evolved during the last decade from domain specific data analytics research. Some of these are WINGS, Taverna, VLAM-G, SciRun, Kepler, XBaya, Vistrails, and Askalon [5][26]. Most of these workflow systems provide GUI interfaces to design workflows and use grid resources to execute them; however, the concepts and structures of workflow components, their representations, data handling strategy, workflow execution engines, and the underlying architectures vary. Comparisons of features of the different workflow systems and taxonomies are presented by Cruz et al. [4], Yu et al. [26], Han et al. [11], and Deelman et al. [4]. While Cruz et al. only focus on a taxonomy for provenance mechanisms, Yu et al. focus on the design, scheduling, fault tolerance and data movement aspects in their taxonomy. Han et al. classify the various types of workflow adaptations, and discuss mechanisms for doing so with reference to the ad-hoc modification requirements for workflows. Deelman et al. discuss a taxonomy specifically for scientific workflows including some of the more recent ones under four main categories, which constitute the workflow life cycle: composition, mapping, execution and provenance. We add a few more high level categories to provide a more general version of the taxonomy and to include some of the key aspects for AaaS on the cloud.

Some of the above workflow systems have been extended to use cloud resources for workflow enactment. Juve et al. [13] discuss the usability of cloud for scientific workflows. Morar et al. [16] present an architecture where Askalon is used to design workflows, which are then executed using the best available cloud resources with respect to costs and SLAs. Ostermann et al. [18] propose the use of a mix of grid and cloud resources when grid resources are unavailable for cost-effective execution of workflows. The benefit is measured in terms of the ratio of the cost of cloud resources and the time saved. We propose hosting the analytics system on the cloud to serve multiple users at different roles using the ubiquitous access, data sharing and scalable features of the cloud.

III. FEATURES OF WORKFLOW SYSTEMS

A. A General Taxonomy

As discussed above, the existing taxonomies focus on a specific subset of the functional properties or features of the analytic workflow systems that prevailed at the time. We provide a more general taxonomy as shown in Fig.1 to include some of the key aspects that we deem are necessary for AaaS on the cloud. The main features of the taxonomy are described below.

1) Structure: Most of the workflow systems are *task-based* where each task represents a data processing or analysis job by a software/service, or a workflow. Back end execution systems such as the Pegasus [4] map the tasks on to grid resources for execution. *Service-based* workflow systems such as Taverna [4], focus on the interfaces for the composition and invocation of services and enable distributed enactment. The information flow can be control, data or a hybrid of the two. Complex workflows include sub-workflows or iterative tasks whereas simple workflows have sequential, parallel or fork type selection structures.

2) Security: For sensitive data, common security methods that are applied are anonymization [6], or encryption [7] including access control measures [1]. Most of the existing systems only have some sort of access control. Anonymization for sensitive data is done before it is used in analysis and needs to be applied to data if it is hosted on the cloud. Researchers are working on encryption mechanisms [7] that would enable the analysis of encrypted data.

3) Workflow Design: Many existing workflow systems such as WINGS, Kepler, Triana, Vistrails and XBaya provide graphical workflow composition tools [5][26]. AVS and SciRun [4] enable users to compose graphical filters and rendering modules to design complex graphics applications. Taverna provides a hierarchical workflow view where a compact high level view can be expanded to see component details. Graphical workflows are typically converted into other representations for storage and execution. Users specify constraints in Resource Description Framework (RDF) format in WINGS, making it a hybrid design method. The constraints allow verification of the workflows. Assistance is provided in the form of suggestions of auxiliary data services, a filtered list of domain-specific software tools and workflow templates, and a mark-up or search functionality.

4) Representation: Workflows can be represented graphically using object-based Unified Modeling Language (UML), Scufl data flow model, graph-based DAX (Extensible Markup Language representation of Directed Acyclic Graph) or petri-net, or event-based BPMN (Business Process Model and Notation) [4], which are easier to construct for small workflows using GUI tools. Text editors are used for parameter specification and descriptions. High level scripting languages such as Ruby [29] and Python [28] are used to automatically generate the low level complex control structures in workflows. High level programming languages are also used instead of GUI tools or text editors to conveniently create the above workflow representations.

5) Visualization: Effective visualization can add an immense value to analytics and can vary based on the resulting data types. Image data should have a good resolution unlike charts or lists. Visualization models can be predefined by the user or defined intelligently by the system based on the data types as in Kepler and VisTrails [4].

6) Collaboration: A data analyst often has limited knowledge about the data domain, which is necessary for designing good analytic models or drawing inferences on

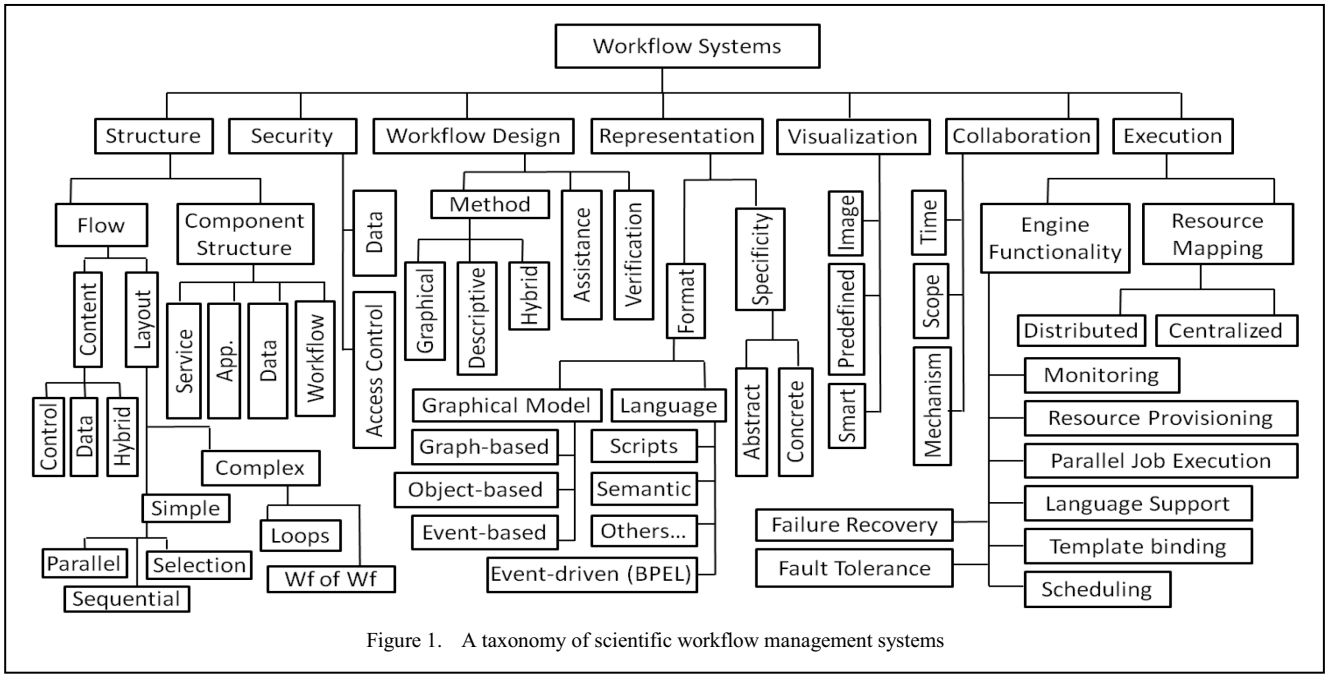

Figure 1. A taxonomy of scientific workflow management systems

analyzed data. Collaboration is, therefore, very important for correct interpretation of the results and specification of effective visualization models. Although the commercial analytic software provide some collaboration functionality, the open source workflow systems have little or no support for online collaboration. Results, data and workflows are shared and published online, for example, in myExperiment [17] and BioMart [2].

7) Execution: Interdependent tasks where the output from one serves as the input for another have to be executed sequentially. Independent tasks can take advantage of distributed parallel execution to maximize resource utilization and minimize the execution time. Pegasus [4][5] takes workflow specifications in XML (Extensible Markup Language) DAG (Directed Acyclic Graph) format and maps them onto grid resources where they are executed using DAGMan [5]. Workflow execution using cloud resources is currently being explored [13][18] due to the scalability required for big data. Depending on the representation of workflows, various workflow engines may be used. WINGS supports multiple engines including shell scripts and Business Process Execution Language (BPEL) [4]. The execution system can be further equipped with monitoring, scheduling, fault tolearance and automated resource provisioning features as in Kepler and WINGS.

TABLE I. CATEGORIZATION OF SOME OF THE OPEN SOURCE SCIENTIFIC WORKFLOW SYSTEMS USING OUR TAXONOMY

	Structure	***Secu-rity***	***Wflow Design***	***Represen-tation***	***Visuali-zation***	***Collabo-ration***	***Execution***
WINGS	Hybrid, app. Data, service, complex	Access control	Hybrid, assistance, verification	Semantic RDF, BPEL, JS, DAGxml	Text, graphviz	-	Dist. / local, multi-format, prov., verification
Kepler	Graphical, app., data, wflow, complex	Access control	Graphical, assistance	Data flow graph	Image, text, graph	Real time, within a group	Dist.,Globus Execution & log, failure mgmt
Taverna	Hybrid, web service based, complex	Access control	Graphical, search/import wflow	XML SCUFL, Freefluo	Image, text, graphviz	Offline, through websites	Centralized Provenance
XBaya	Hybrid, web service based	-	Graphical, assistance	Jython script	Text	-	Script, centralized
Triana	Hybrid, web service	-	Hybrid, wizard, verification	WSRF, web services	Graph, image, text	-	Centralized
VisTrails	Hybrid, app., data, simple	Access control	Hybrid with versions	XML with annota-tion	Image, text	Offline, shared DBMS	Centralized, script, multithread

B. *Required Features for an AaaS*

We present a general taxonomy of workflow systems based on 7 main features, which are used in Table 1 to categorize some of the popular open source scientific workflow management research tools. We explore the different features up to various depths depending on the focus of our research. Based on our study, an analytic workflow system should have the following key features:

- Hierarchical composition of workflows as shown in Fig. 2. The hierarchies include data schema and metadata definition and pre-processing, specification of analytical model(s), software or service configuration (with analytical models, parameters and data links) and verification, and workflow composition (using the above or another workflow). Each level in Fig. 2 represents a small workflow as a revision may be needed based on expert feedback until the results are satisfactory. Also for simple analytic jobs where a simple analytic query or a custom application can be used, level 2 may be skipped to move to level 3 from level 1. Moving back to level 1 from level 4 allows modification of pre-defined workflows and re-use of pre-defined analytic models and workflows in new workflows. Therefore, the key features of workflow systems should include:
  - Support for templates and versioning to enable revision and reuse of templates in multiple workflows
  - Support for dynamic design and validation of workflow components before composition
- Support for sequential, parallel, iterative and selective flows
- Support for data and/or control flow
- Specification of constraints and dependencies for verification and optimal parallel execution of workflows
- Quality of Service (QoS) provisioning based on SLA for Analytics-as-a-Service
- Transparent use of scalable cloud resources for cost-effective execution of workflows
- Support for provenance with effective logging
- Support for a recommender system to provide domain specific assistance in the design of analytic

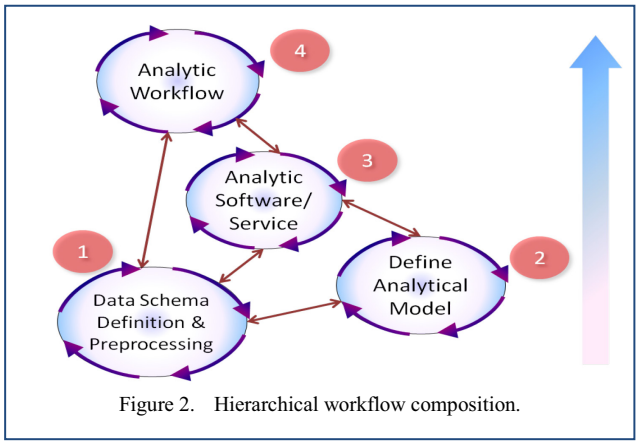

Figure 2. Hierarchical workflow composition.

and visualization models and execution of workflows
- Focus on user role, context, and business aspects in providing customized GUI and services
- Expandable graphical view of high level compact workflows
- Minimum data movement
- Data security and privacy

C. User Groups

An AaaS should provide custom interfaces for different user groups or roles, which would typically include:

1) Scientific Analysts - Have knowledge about analytical tools and methodologies, may or may not have knowledge about the data domain, e.g. statistician/data mining experts. Requires access to most of the functionality provided by CLAaaS including software tool definition/import, workflow template design, scheduling, and execution, visualization, and collaboration.

2) Domain Experts - Know the data very well and understand the implications of the various data values, usually guide the analyst and help to draw inferences from the results, e.g. medical doctors, meteorologists. Requires access to functionality such as pre-defined template-based simple workflow definition by data binding and execution, collaboration, visualization and analytic requirement specification.

3) Practitioners - Have partial understanding of the data values and run pre-defined workflows as end users, e.g. nurse practitioners. Require access to functionality such as execution of simple workflows with new or updated data, query analyzed results, visualization and collaboration.

4) Administrators - Make secondary inferences to derive business values, and thereby, take administrative actions, e.g. sales or staff administrator. Require access to functionality such as information requirement specification, query analyzed results of specific data, visualization and collaboration.

5) Managers - Make long term strategic planning based on corporate goals, and accordingly request for specific information or insights on various aspects, e.g. CEOs. Require access to functionality such as information requirement specification which may call for predictive analytics or multi-source data analytics, query analyzed results of specific data, visualization and collaboration.

A user can have multiple roles and belong to multiple data domains. An AaaS should, therefore, show customized GUIs for different domains and user roles. The SLA will dictate the various user groups and the corresponding access policies of a registered service consumer, which will typically be an organization of multiple individuals. Once the SLA is signed, users belonging to the organization will individually register against the corporate SLA with their role and other personal information.

IV. CLAaaS

A. CLAaaS Architecture

We propose CLAaaS, a cloud-based AaaS provisioning platform, which would address the shortcomings of the existing systems and include the required features as discussed in Section 3. Fig. 3 shows the conceptual architecture of CLAaaS. As for any cloud service, the internet traffic passes through firewalls and is directed and distributed via web proxy and load balancers. The main parts of CLAaaS are the set of front-end portals and the core computing platform hosting the back end components. The compute and storage clusters on the right hand side imply scalable cloud resources that are used by CLAaaS as needed. We describe the components of CLAaaS below under three functional categories: *Service Management*, *Workflow Management*, and *Data Management*. The modular architecture of CLAaaS allows specific components to be replicated for scalability or proximity to data.

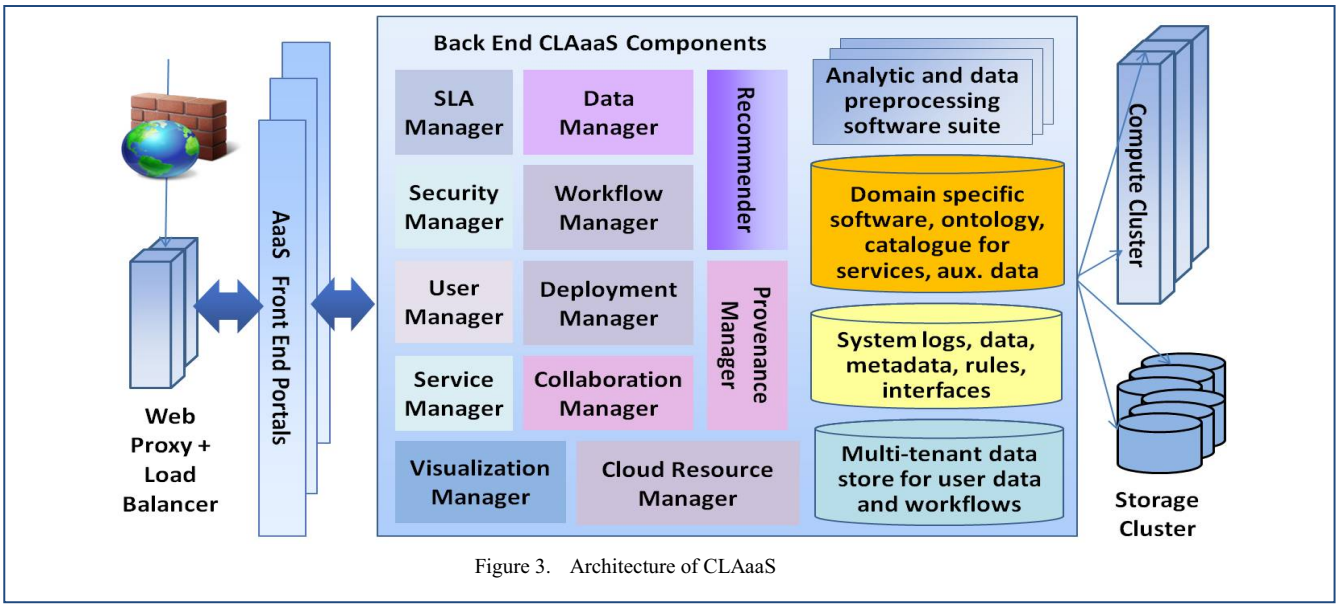

Figure 3. Architecture of CLAaaS

Usability, customizability and SLAs are the key components of any service provisioning system. The *AaaS front end* comprises a set of web portals customized for different data domains and user groups, which provide role-based controlled access to CLAaaS for user registration, collaboration, data query, and workflow design and execution. Therefore, they encompass all three functional categories. The rest of CLAaaS back end components on the core platform are described below.

1) Service Management Components: The SLA Manager, Security Manager, User Manager, and Service Manager perform the service management tasks. The Service manager is responsible for the overall status and functionality of the system. It oversees the requests, maintains session information, and coordinates the functions of the different components by communicating with them as necessary. The SLA manager handles negotiation of SLAs for the different levels of services and user groups and monitors the compliance of the SLAs. The User manager helps in user registration, maintenance of user accounts, and the definition and authorization of user groups. Security of the system is ensured by the Security manager according to the pre-defined system access policies and those specified for the user groups.

2) Workflow Management Components: The Data Manager, Workflow Manager, Deployment Manager, Visualizaton Manager, Collaboration Manager, Recommender, Provenance Manager, and the Cloud Resource Manager together constitute the workflow management system. The Workflow manager helps to create a new workflow or to modify existing workflows. The Deployment manager is responsible for the mapping of the workflow on cloud resources. It then executes workflows transparently using cloud resources provisioned by the Cloud Resource manager, and handles failures and recovery. The Data manager helps users in managing, importing and linking data sources, and specifying data dictionaries, metadata and ontology. Assistance for the design and deployment of workflows for different data domains is provided by the Recommender. The Provenance manager tracks the parameter values and steps during a workflow design process and monitors execution trails. This information is used by the Recommender to create recommendations and the Deployment manager to detect and recover from failures. Visualization manager helps users to design templates for viewing data and results of the analysis in a preferred format, which is particularly important for high resolution image data. Finally, the Collaboration manager allows offline messaging or live collaboration among members of predefined user groups.

3) Data Management Components: The analytic and data processing software suite and the three data stores on CLAaaS are grouped under this category as they all perform data management or processing tasks. The software suite can consist of proprietary and custom data processing or query application software, including external data or software services. It can be further customized by the platform users as needed. The three main data stores on the platform contain:

a) *Domain specific data*: Includes commonly used domain-specific analytic tools, links to services, workflow templates, ontology, and data sources, which are catalogued to facilitate query and re-use.
b) *System specific data*: Includes general system related data, metadata, list of cloud resources and concurrent users, pre-defined rules and policies, logs, and portals for the service operation.
c) *User specific data*: Includes personal data of the users registered under a corporate or individual SLA, uploaded or linked data for analysis, analyzed data and visualization models, and analytic models and workflows.

B. Technology and Concepts for Implementation

Our initial implementation of CLAaaS platform will be in the Virtual Computing Lab (VCL) cloud environment [24]. The analytic software suite will primarily include SPSS, Cognos, InfoSphere BigInsights and InfoStreams from IBM [12]. SPSS is widely used in predictive analytics and offers a rich set of statistical and data mining algorithms which are used to define analytic models. Cognos is the IBM business intelligence software and offers rich collaboration and reporting features. InfoSphere BigInsights is built on the Hadoop framework and offers the power of parallel computation for big data and supports Apache data analytics tools [10]. InfoStream enables processing of large data streams [12]. Together they provide a rich toolset for data analytics. We would also like to include popular open-source tools such as R [19] and Weka [25] in the software suite. We plan to adapt one of the open source scientific analytic workflow systems depending on its support for the necessary features and available example data and workflows as in myExperiment [17] and BioMart [2]. The adaptation will be necessary to support multi-tenant data management and scalable mapping of the workflows on the cloud virtual machines (VM) for execution.

Multi-tenancy is a commonly used technique in the cloud environment where multiple customers share a single software or hardware resource as tenants. It allows cost effective on-demand use of resources with some compromise in the design and implementation effort, performance, and security. A *multi-tenant software* provides simultaneous access to multiple users or tenants transparently through customized interfaces. A *multi-tenant data store* maintains data from multiple users in the same storage system for cost-effective use of compute, memory, and storage resources. Different approaches are used by both multi-tenant software and data stores to keep each user unaware of the existence of the other users, and to protect security and privacy of users' data. For software, separate servers ensure maximum separation of users' data but the most cost-effective approach applies multi-threaded approach on a shared VM. However, in the latter case the software must be designed to ensure separation of the data of different users. Otherwise, separate application servers can be used for the users on a shared VM, or separate VMs can be used to run multiple instances of the application for different users. We will adopt the multiple

instance approach to multi-tenancy for the single user software tools.

For data store, maximum user data isolation is guaranteed by using separate physical storage servers for different users. More cost effective solutions in the cloud apply one of the following approaches listed in the order of decreasing isolation of user data: a) a single server hosting multiple VMs, each running a separate storage system for a user; b) a server with a storage system hosting multiple databases, one for each user; c) a server with a storage system containing a shared database where each user has a separate schema, and d) a server with a storage system containing a database with a schema that is shared by multiple users, and the user-specific keys are used to identify the corresponding rows of data. We intend to apply the last one, shared database and shared schema, approach to multi-tenancy for the system and some user specific data such as user profile, user-specific software configurations, and system usage. For the user data that need to be analyzed, we plan to use the shared database and separate schema, or the separate database approach to multi-tenancy as applicable based on the variations in data types.

CLAaaS will be accessible through the front-end services and web portals running on one or more VMs. We will configure a number of VM images with an interface to CLAaaS platform and likely combinations of the software. To support specific level of users and running workflows, the platform VM will configure and instantiate other VMs as necessary with the appropriate machine image [15][18].

C. A Use Case Scenario

Analytics is used widely in today's digital world. In the case of calamities or emergency situations, CLAaaS can be of particular help where users do not have to go through system setup and can immediately launch the required analytic platform on the cloud. For example, in the case of a sudden deadly attack at a national event, analyzing the data about relevant factors such as the organizers, the guests, collocated events, the venue, attack strategy, and used firearms can reveal important information about the possible suspects and their motivations. CLAaaS can be used to integrate and process data from multiple sources such as police, government sources, hospitals, news sources and the internet. Anonymization techniques [6] can be applied to sensitive personal data if necessary prior to uploading it on the cloud, or CLAaaS can also be implemented on a private cloud. The integrated data can be made accessible to experts such as police chiefs and government intelligence. Analysts will use CLAaaS, define analytic models and workflows to discover further information with guidance from the experts, and high officials. Selected analyzed information can be published for the guest reporters and other concerned groups as needed. Analytic models providing the most effective information can be saved for re-use. For a data source containing sensitive data that cannot be uploaded on the cloud, external data services can be invoked by the workflow management system in CLAaaS to execute the requested analysis externally and retrieve the results such as information about foreign citizens.

The experts, analysts, and executive officers can use CLAaaS collaboration tool to discuss the findings, request for further information and annotate important observations. As new information about the suspects are uncovered, further analysis is called for to find probable accomplices, which results in a multi-step workflow. The workflow template can be saved to analyze similar data sources in the future. The high commanders can access CLAaaS to receive updates, communicate with specific groups, issue orders, or request for special reports.

On the service provider's end, the investigating office can be the main customer to sign the SLA, which will ask its designated personnel to register with CLAaaS in different roles. CLAaaS will be configured with necessary big data analytic software such as BigInsights, collaboration and reporting software such as Cognos, and domain specific data such as police offices, city, and postal codes. Each registered user can then log in to carry out individual tasks in the designated roles, i.e., data analyst, domain expert, guest reporter, executive officer, and the high commander. After the job is done, the service can be unsubscribed but the analytic workflow templates and other relevant data can be downloaded in a persistent storage for future use.

V. CONCLUSION

Analytics is crucial in providing insights on big data for efficient decision making. Analytic workflows compose the various data processing and analysis steps needed to extract values out of the data. Researchers have developed multiple workflow systems [4][26] during the last decade, which were initially geared towards specific data domains and had different design objectives. Several industrial [12][20][21] and open source [19][20] solutions exist as well. In this paper, we present a taxonomy to identify important features of analytic workflow systems, possible user groups, and propose CLAaaS, a cloud-based AaaS provisioning platform.

Researchers are already sharing some data and workflows on the cloud [2][17]. We intend to leverage this further by provisioning Analytics-as-a-Service on the cloud using CLAaaS platform infrastructure as shown in Fig. 3. It would a) implement multi-tenancy for a wide range of analytic software tools and back end data sources; b) provide SLAs and customized interfaces for different groups of users; c) enable scalable data management and workflow execution for big data, and d) promote web collaboration. Concerns for data privacy in the cloud [3] can be addressed by implementing CLAaaS in a private cloud, applying security mechanisms [7], and distributed enactment of the workflows using 3rd party services [4]. Considering the huge amount of data that exists on the web today, CLAaaS would lead the way with a scalable ubiquitous access to the power of analytics.

ACKNOWLEDGMENT

The research is supported by the IBM Canada Research and Development Center and the consortium of Southern Ontario Smart Cloud Innovation Platform (SOSCIP).

REFERENCES

[1] Alharkan, T., and Martin, P., 2012. IDSaaS: Intrusion Detection System as a Service in Public Clouds. In *proceedings of International Conference on Clouds, Grids and Virtualization*, Nice France.

[2] BioMart software and data services, available March 5, 2013 at: http://www.biomart.org/.

[3] Chen, D., and Zhao, H., 2012. Data security and privacy protection issues in cloud computing. In *proceedings of International Conference on Computer Science and Electronics Engineering (ICCSEE), Vol. 1, pp. 647-651,* IEEE.

[4] Cruz, S., Campos, M., and Mattoso, M., 2009. Towards a Taxonomy of Provenance in Scientific Workflow Management Systems, *in proceedings of IEEE Congress on Services– I*, Los Angeles, CA, USA.

[5] Deelman, E., Gannon, D., Shields, M., Taylor, I., 2009. Workflows and e-Science. An overview of workflow system features and capabilities, *Future Generation Computer Systems, Elsevier, Vol. 25(5)*, *pp. 528-540*.

[6] Enhancing Cloud Security using Data Anonymization. IT@Intel White Paper 2012. Available Feb. 28, 2013 at: http://software.intel.com/sites/billboard/sites/default/files/downloads/enhancing-cloud-security-using-data-anonymization.pdf.

[7] Gentry, C., and Halevi, S., 2011. Implementing Gentry's fully-homomorphic encryption scheme. In *proceeding of Advances in Cryptology–EUROCRYPT'11, pp. 129-148*, Tallinn, Estonia.

[8] Gil, Y.; Ratnakar, V., and Fritz, C., 2010. Assisting Scientists with Complex Data Analysis Tasks through Semantic Workflows. *In proceedings of the AAAI Fall Symposium on Proactive Agent Assistants*, Arlington, VA, ACM.

[9] Google Analytics. At http://www.google.ca/analytics/.

[10] Hadoop. At: http://hadoop.apache.org/.

[11] Han, Y., Sheth, A., and Bussler, C., 1998. A taxonomy of adaptive workflow management. In *proceedings of the Conference on Computer Supported Cooperative Work (CSCW-98)*, Seattle, WA.

[12] Howard, P., 2011. IBM and big data: an introduction. *IBM White paper*. Available March 6, 2013 at: http://www.bloorresearch.com/analysis/11737/ibm-big-data-introduction.html.

[13] Juve, G., Deelman, E., 2010. Scientific workflows and clouds. *Crossroads*, *Vol. 16(3), pp. 14-18*, ACM.

[14] Kim, H., Cho, I., and Yeom, H., 2008. A Task Pipelining Framework for e-Science Workflow Management Systems. In *IEEE International Symposium on Cluster Computing and the Grid, (CCGRID'08)* pp. 657-662, IEEE.

[15] Mian, R., Martin, P., and Vzquez-Poletti, J., 2012. Provisioning Data Analytic Workloads in a Cloud. In Press for *Future Generation Computer Systems*.

[16] Morar, G., Muntean, C., and Silaghi, G., 2011. Implementing and Running a Workflow Application on Cloud Resources. *Informatica Economica*, *Vol. 15(3), pp. 15-27.*

[17] myExperiment workflow hosting site: available March 5, 2013 at: http://www.myexperiment.org/home.

[18] Ostermann, S., Prodan, R., and Fahringer, T., 2010. Dynamic Cloud provisioning for scientific Grid workflows. In *proceedings of IEEE/ACM International Conference on Grid Computing*, *pp. 97-104*, IEEE.

[19] R-Project. At http://www.r-project.org/.

[20] RapidMiner from Rapid-I. Available Feb. 26, 2013 online at: http://rapid-i.com/content/view/181/190/.

[21] SAS. At: http://www.sas.com/.

[22] Slot, M., and van Zoolingen, P., 2005. Workflow Management Systems. *Technical report*, Div. of Math. and Computer Science, Vrije Univ., The Nerherlands.

[23] Tableau. At: http://www.tableausoftware.com/.

[24] The VCL Cookbook, 2010. *IBM Global Education*, white paper at: ftp://ftp.software.ibm.com/common/ssi/ecm/en/ebw03004usen/EBW03004USEN.PDF.

[25] Weka. At: http://www.cs.waikato.ac.nz/ml/weka/.

[26] Yu, J., and Buyya, R., 2005. A taxonomy of scientific workflow systems for grid computing. *Sigmod Record, Vol. 34(3)*, *pp. 44-50*.

[27] Zarandioon, S., Yao, D., and Ganapathy, V., 2012. K2C: Cryptographic cloud storage with lazy revocation and anonymous access. *Security and Privacy in Communication Networks, Vol. 96, pp. 59-76.*

[28] Python programming. Available May 1, 2013 at: http://www.python.org.

[29] Ruby programming: Available May 1, 2013 at: http://www.ruby-lang.org/en/.