In-Memory Modeling and Analytics as-a Service in Cloud

Prabal Mahanta and Dhwanit Shah
SAP Labs India Pvt. Ltd., Bangalore, India
Email: {p.mahanta, dhwanit.shah}@sap.com

***Abstract*—The aim is to conceptualize and create a platform for building complex data models and executing huge data processing tasks in a more efficient manner. We try to conceptualize a solution which caters to cloud infrastructure and memory-resident storage platform by providing a virtualized environment to end-users for carrying analytics on large scale without investing huge cost from infrastructure setup and maintenance point of view. The solution will support an analyst to replicate data and compute-intensive job running on his/her personal machine, to run on cloud with minimal effort and providing scope for future scale-out. There are several applications which are intensive in terms of data storage and computation cycles. For such applications hardware infrastructure and maintenance costs considerably increase with the increase in data volumes. These applications will vastly benefit with large scale in-memory database management system in cloud where the computations can be done with minimal effort. The idea is to enable provisioning for end users to custom build the skeleton of their data models for data manipulation and then export them using analytics service on the cloud. The ability of the system to expose results of computation in various data format will enable different unique insights to the data. The platform will enable users to harness the power of cloud and in-memory database to solve complex real-time problem with minimal effort and reduced cost.**

***Index Terms*—In-Memory Database, In-Memory Compute Engine, Cloud Service, R, Amazon EC2, SAP HANA, Analytics**

I. Introduction

Business intelligence and real time analytics enables a business to foresee the expectations from a consumer's perspective and hence providing a scope for better planning and decision making for complex business scenarios. These complex business scenarios require complicated analytical models to be generated. Using these models it is required to process huge sets of data. The goal is to conceptualize a platform for analyst who will be able to build complex data models and process data in a secure and effective manner.

Virtualized infrastructure for end-users has evolved the cloud computing domain where the users utilize the computing power for processing their tasks and produce detailed insights to the business scenarios. Also leveraging memory-resident storage and compute engine, enables the business applications to be more responsive as required by today's enterprise applications like banking, supply chain and scientific applications (for example - classification of gene expression data [1], deployment and execution of scientific workflows like brain imaging workflow execution [2], Geographic Information Systems (GIS) [3], oceanic and climatic simulations, analysis and predictions [4]). The common feature among these applications is the large feature space/dimensionality in the data set and varying workloads.

We propose a hybrid solution which enables an analyst to replicate a data intensive job for computing data sets of the order of terabytes with minimum effort and resources. The concept will help business analyst to securely move the data from his/her personal workstation over the cloud. The user can then utilize the offerings in the cloud - in-memory storage and compute engine. The idea is to have skeleton models where user can custom build their models for data manipulation and then gracefully expose them to the cloud over a secure channel in an encrypted mode with the help of cloud service. The application is executed by in-memory compute engine in cloud environment and result is exposed to the client for consumption in standardized data-interchange format. Thus an analyst having limited technical development abilities but strong statistical and analytical skills can harness the power of cloud and in-memory storage platform to solve complex modeling and real time analytical problems in very little time and money.

II. Related Work and Challenges

There are several cloud vendors who provide variety of solutions in all the three major layers of cloud computing namely Infrastructure as-a Service (IaaS), Platform as-a Service (PaaS) and Software as-a Service (SaaS). For e.g. CloudFoundry [5], StarCluster [6], RightScale [7], Heroku [8] and many others are cloud computing service vendors focus on providing highly configurable compute clusters in cloud operated infrastructures for use by highly skilled and technically sound programmers and developers. But our idea focuses on driving cloud computing solutions for users who possess limited technical capabilities in terms of setting up cloud infrastructure, creating and deploying instances, configuring environment, creating complex models for

Manuscript received July 18, 2013; revised September 15, 2013.

doi: 10.12720/jiii.1.4.211-217

data deployment and coding complex application logic for performing analytical jobs. Apart from that existing technical challenges faced, a business analyst has to deal with loading large amount of data from various resources available in various forms in to the cloud, deal with risk involved in securely consuming cloud services, transferring data to cloud and accessing data from cloud. Additionally, cloud computing solutions require analyst has to learn the architecture and process of data generation, data transmission, data storage and data retrieval in cloud infrastructure. Above mentioned challenges forces a business analyst to understand the working of cloud computing, service models, deployment architecture, application development and risk mitigation techniques, consuming his/her valuable time and money which he/she could have possibly spent on productive analysis and decision making tasks.

III. Our Approach

A. New Service Model

In this paper we demonstrate a system level architecture of enabling cloud infrastructure to provide custom data modeling and analytics as a service. Fig. 1 demonstrates a conceptual design of general cloud computing services available at various layers of cloud infrastructure offerings often referred as service model.

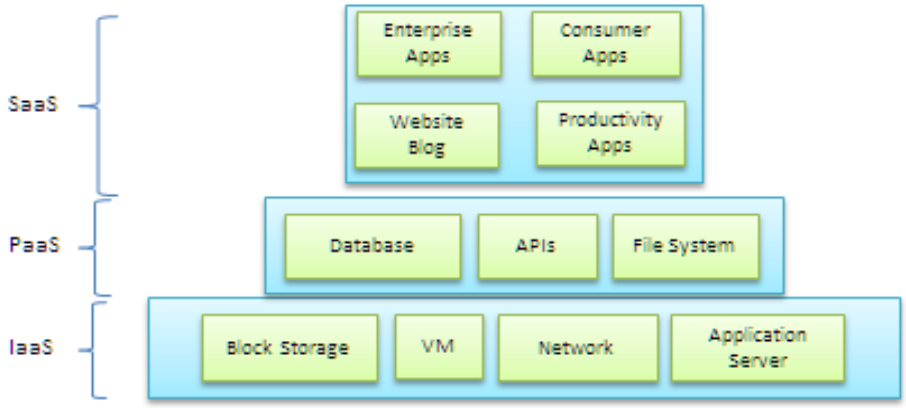

Figure 1. Cloud Service Architecture

IaaS is the hardware and the software that runs the cloud infrastructure at the lowest level in a cloud computing stack. IaaS consists of the storage, application servers, networks and the operating system virtualization. PaaS forms the second layer in cloud computing stack. It provides all the tools and services required for a developer to quickly code and deploy applications on the top of IaaS layer without caring about the underlying infrastructure. The topmost layer is the SaaS layer in the cloud computing stack. This layer provides applications designed for the end users and are mostly delivered over the web. SaaS provides services at enterprise level targeting business services and mission critical apps as well as productivity apps like office automation, email and content sharing. Consumer centric apps like wikis and blogs are also available in SaaS layer. The service models keep on changing based on what fits best based on current user's requirements.

Our solution utilizes all the three service layers of the cloud computing stack and creates a new service model which tries to mitigate the risks and challenges faced by any user, be it an auditor or an analyst. We call this in-memory modeling and analytics as a service in cloud or IMMAAASC. In this paper, to demonstrate our idea, we have considered a person called analyst who is a typical user of our newly proposed service model. The analyst wants to create custom models from the large data set and perform multivariate analysis on the top of it in order to find patterns or solve optimization problems. Here for conceptualizing this scenario we have tried to understand the analyst's problems in terms of technical expertise and domain level issues. Every domain has a specific requirement and customized models will help perform efficient analytics and optimize the data set representations so that with minimal technical effort, we are able to cater to the needs of the analyst and provide a solution where data management and analytics is optimized for enabling ease of use.

Our solution is divided into two sections namely the conceptual design of IMMAAASC service layer and secondly how this new layer fits in to the existing cloud computing stack enabling an analyst to get his/her job done in a manner which is extremely fast, easy and secure. There are primarily two kinds of cloud services available namely public cloud and private cloud. Public cloud is generally owned by an enterprise provider who offers an easy to set up and inexpensive cloud solution to the general public thereby leveraging the high end cloud computing capabilities towards the growing business demands of the customers. Being flexible in terms of usage, it allows the enterprise users to focus on addressing business challenges rather than dealing with the scalability and infrastructure issues. On the other hand, private cloud is meant to be adopted by enterprises requiring more administrative control and risk mitigation in terms of cloud deployment and its usage. IMMAAASC can be

typically deployed in both scenarios. It brings about the best of both worlds by giving the end user the ability to massively scale the application while providing enhanced security and reducing deployment risks.

B. IMMAAASC Architecture

There are several key factors that influence the performance and capabilities of modeling and analysis of an enterprise cloud providing analytics as a service:

1. Number of daily inflow of new data records.
2. Volume maintenance.
3. Structuring unstructured data.
4. Analyzing data structures that can provide insights

The impact of big data segment and the credibility of insights from the analysis present a challenge to domains like finance and life sciences to intelligently manage data. The service layer will provide architecture to mitigate the challenges in infrastructure building and maintenance by providing standardized method of managing structured and un-structured data which can be seen in Fig. 2a. The data is ported to the ecosystem of IMMAAASC through a standardized cloud gateway using secure upload service. After data migration, the in-memory computing engine helps in transforming the data in a structured manner and this service can be applied to the incoming data stream and this helps to process the data on the fly and apply the transformation The storage service extracts and parses the metadata to build an in-memory schema which can be later utilized in future data load. There exists a choice of selecting storage mechanism which can either be column store or row store. As shown in Fig. 2b, our service will be exposed for the analyst to perform delta load operations of business processes data in a secure manner over an encrypted channel.

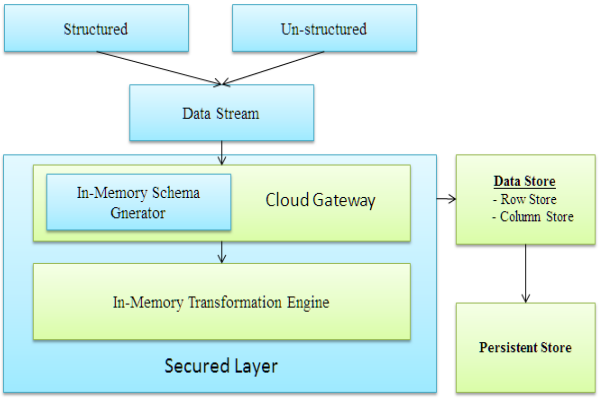

Figure 2a. Data transformation and storage

Here we consider two scenarios where an analyst can perform analysis on the fly. As shown in Fig. 2c, in-memory compute engine is installed and configured to work in integration with R [9] on an Amazon Web Services infrastructure consisting of Amazon Elastic Compute Cloud (EC2) [10], Amazon Elastic Block Storage (EBS) [11] and Amazon Simple Storage Service (S3) [12]. In-memory database forms the basis for handling storage of enterprise and public data streams in a dynamic way by building models and schemas on the fly. We have used HANA [13], SAP's in-memory database platform available over the AWS cloud called SAP HANA One [14]. The decision of storage orientation for keeping data in the database, whether column oriented or row oriented is left to user's discretion. For our purposes we favor storing data in an in-memory column store for faster data retrieval in OLAP scenarios. The dynamic schema mapper is an important part of the IMMAAASC service platform because it is responsible for decoupling the complexities involved in implementing mechanisms for management of data like modeling, storage, retrieval and security, from the actual purpose of an analyst i.e., analysis of data.

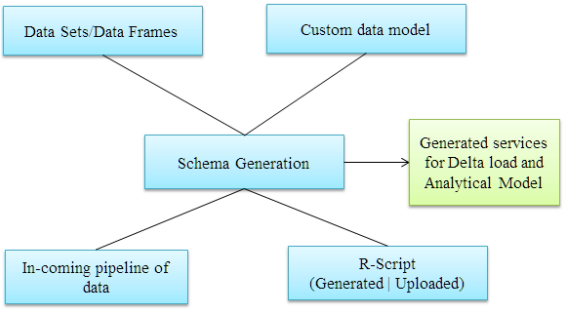

Figure 2b. Custom data and model mapper with schema generation

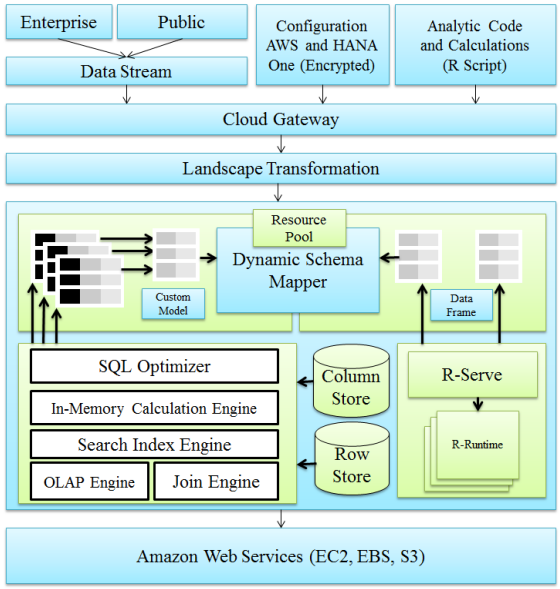

Figure 2c. High level architecture

The architecture being tightly coupled with the R engine, the scenario for study is, when a user uploads flat files in the phase of data load and a schema is generated using the mode of custom development where the user merges the self-designed model with the system generated ones. Once the schema is generated the database table is then mapped with the model based on the fields marked as required by the models. At a high level, the schema mapper is responsible to map the data from an in-memory engine to the R data frames and data tables residing in the memory pool shared by R-Runtime.

C. Dynamic Schema Mapper

As shown in the Fig. 3, the dynamic schema mapper takes an R script created by the analyst for analysis as input and parses the script to uniquely identify and mark the resources required for the logical computations performed in the calculations. We have used xml format to store the metadata containing schema, tables and field corresponding to the incoming data stream. At first, when the R script is uploaded to the cloud, an in-memory parser creates an xml markup map by mapping unique identifiers for each of the data frames and data tables used in the mathematical model to the corresponding tables and fields present in the schema in an in-memory database. Once schema is generated for the incoming flat data files, together with the xml markup containing the mapping for every data frame and data table to the corresponding database field and table, the dynamic schema mapper creates code for modeling the R variables using standard SQL statements and their result sets. Thus a mathematical model is created in which the corresponding data frames and data tables in R script are integrated with the model created in the database schema. In this way IMMAAASC in-memory parser and mapper will parse and map the R code respectively, and builds the mathematical model. This mathematical model is mapped to the R-Runtime libraries integrated in the in-memory engine. Moreover the mathematical model can be replicated and used as prototype for other users to create model and perform analysis in cloud. In a typical analysis scenario, an analyst securely transfers data and mathematical calculations involving fields of a table in an R script to the cloud environment using the IMMAAASC service model. In another analysis scenario, considering data is already present in the in-memory database in the cloud and the analyst wants to create mathematical models for analysis on the top of the data. IMMAAASC platform can facilitate the requirement by providing a user interface where the analyst can simply select, drag and drop the table fields and enter formulas and quantitative calculations using the selected fields, and the IMMAAASC dynamic schema mapper creates the underlying SQL code for the task on the fly thereby creating a complete mathematical model for analysis. Here also the mathematical model will be mapped to the R-Runtime libraries in the in-memory engine. In this scenario, the time to design and preview will drastically come down when template based modeling can be mapped to this case. The following diagram (Fig. 4) presents the high level design of the IMMAAASC parser and mapper engine. The output of the analytical job is stored in the result tables in a relational format for the computations returning data set containing multiple records. Analyst can also choose to store the result in output files generated in the IMMAAASC workspace. Also the scalar outcomes are stored in scalar type cell uniquely identified by the analytical job run identifier. The result can be consumed by REST services or output files can be downloaded via a secure channel requiring user and application authentication before access. Analyst can generate the report from the data stream output of the service.

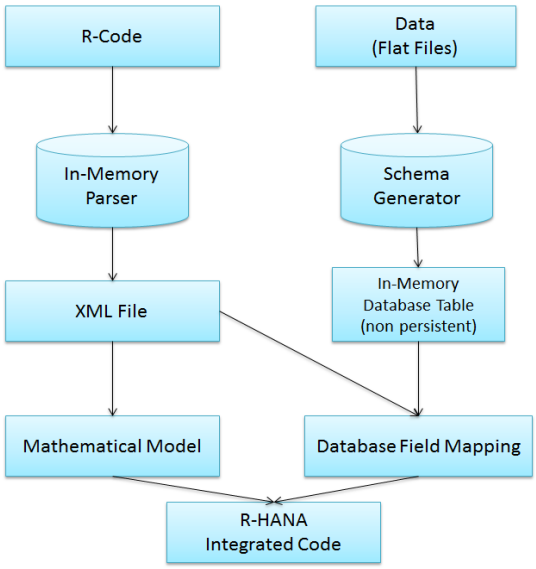

Figure 3. Dynamic schema mapper workflow

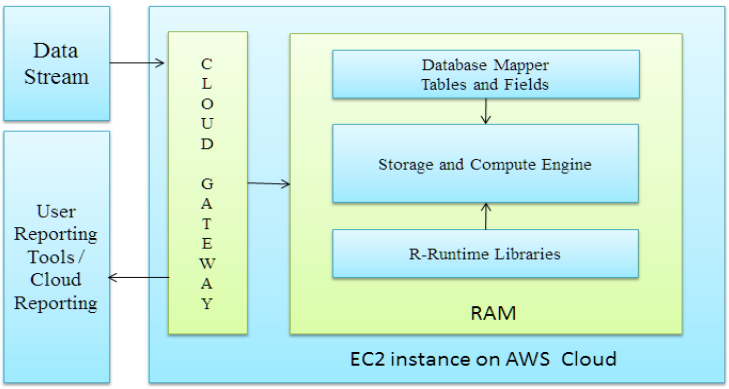

Figure 4. Service data manager

IV. Implementation

IMMAAASC is another service layer between the analyst personal workstation and the underlying cloud infrastructure. In our case, the IaaS is provided by

Amazon Web Service (AWS). IMMAAASC platform uses the boto python package [15] to provide connections and interfaces to the EC2 and EBS remotely using the AWS service. An analyst having an AWS account can create EC2 instance on the AWS using the IMMAAASC platform and can even mount an EBS volume to the newly created EC2 instance. EBS provides permanent block storage to persist the data on the cloud. Using this interface, EC2 instances can be launched, stopped or terminated. Once an EC2 instance is created and launched, we need to assign an Amazon EC2 Elastic IP Address (EIP) [16] to the AWS instance. After procuring EC2 instance, the analyst has to launch SAP HANA on the newly created instance by attaching it to one of the available AMI (Amazon Machine Images) [17]. This can be felicitated by navigating to the SAP HANA One management console on AWS website. By launching SAP HANA on an EC2 instance, we have configured and launched an in-memory storage and compute engine. The IMMAAASC platform requires a configuration file for storing the EIP (Elastic IP), the host to connect on AWS, SAP HANA Database username and password. The configuration file is encrypted for security purposes.

For e.g. xxx.xxx.xxx.xxx hanaserver hanaserver.compute-1.amazonaws.com (EIP is masked in this paper), is host on AWS.

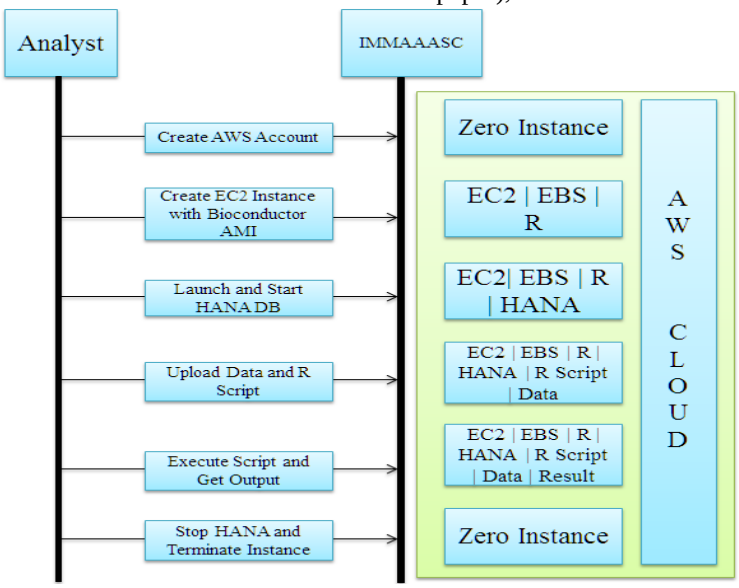

Figure 5. Typical IMMAAASC workflow

The EC2 instance(s) are initialized using the Bioconductor Cloud AMI. The live AMI ids are available on the Bioconductor AMI launch and configuration page [18]. These AMI ids are provided in the configuration file. The RServe package [19] is preloaded with the Bioconductor AMI. There are other R packages that IMMAAASC has to install and configure once the HANA Database are launched and configured on an EC2 instance. HANA can be connected from an R-Runtime using the RJDBC [20] package or RHANA package. Once the core packages are installed in R, an IMMAAASC workspace is initialized for the current user session on the cloud instance. The workspace is used for storing data files, computational R scripts along with the output result files. The R script containing the mathematical calculations and the flat data files are uploaded to the repository of IMMAAASC created at the EC2 instance in a secure way by either using SCP (Secure Copy) protocol or rsync protocol. Once the data and script are uploaded to the IMMAAASC repository, the transform engine creates a database schema in the HANA database, cleans the data, parses the data, creates appropriate tables and Variable Character Field columns, imports the data from the data files into the database tables and generates an XML mapping file that uniquely identifies each field in each table from the schema into the uploaded data set. On the other hand, the dynamic schema mapper parses the R script to create appropriate mappings of the data frames and data tables into the database schema. An analyst can also make necessary manual fixes if the mapping is invalid. After the correct mapping is identified and committed, the R script is updated to include database calls using RJDBC or RHANA that creates a dynamic model by fetching the corresponding multidimensional data from the database schema and make them available for use in the R-Runtime environment when the script is executed. At the end of the analysis session, analyst can choose to either store the output in the file on the IMMAAASC workspace or store it in the database table of the selected schema. The output can be downloaded by using SCP or rsync or fetched from the database using REST service. During the processing of an analysis script, analyst can also see the status of current execution process by querying the status table in the database corresponding to selected analytical run id. On the completion of the analysis task the user can choose to stop the HANA DB instance and even EC2 instance. Fig. 5 shows the work flow of steps starting from the procurement of HANA on AWS through the entire analysis process. The platform [21] for IMMAAASC is also configured for batch mode execution where the configuration parameters are specified fully with the data and code pre-loaded in the

cloud site. IMMAAASC service model is also effective in terms of usage and service cost. The cost of moving the data to and from EBS volumes can be greatly reduced because all the data for analysis resides in memory in the HANA database tables. By configuring the data storage orientation and HANA performance parameters, one can make optimal use of the underlying CPU resources allocated and boost the performance of aggregate queries, OLAP joins, Top-N analysis, Group By, Sort queries etc.

V. Impact

All the end users, business analyst and even the starts-ups customers require a platform where they can plug and play with their data to gain valuable insights in an intelligent manner [22]. Visualizing the output in a convenient tool using the service oriented framework will change the way analytics domain are in place in present times.

Fig. 6 shows a typical client-server set up for the IMMAAASC platform and here user can also simulate the real-time data over the web server [23] connected to the HANA database on AWS [24] cloud. By using sophisticated UI tools and interfaces on the client side, the end user can directly carry out various analytics scenarios on real time data using high speed in-memory compute engine running on high performance cloud infrastructure.

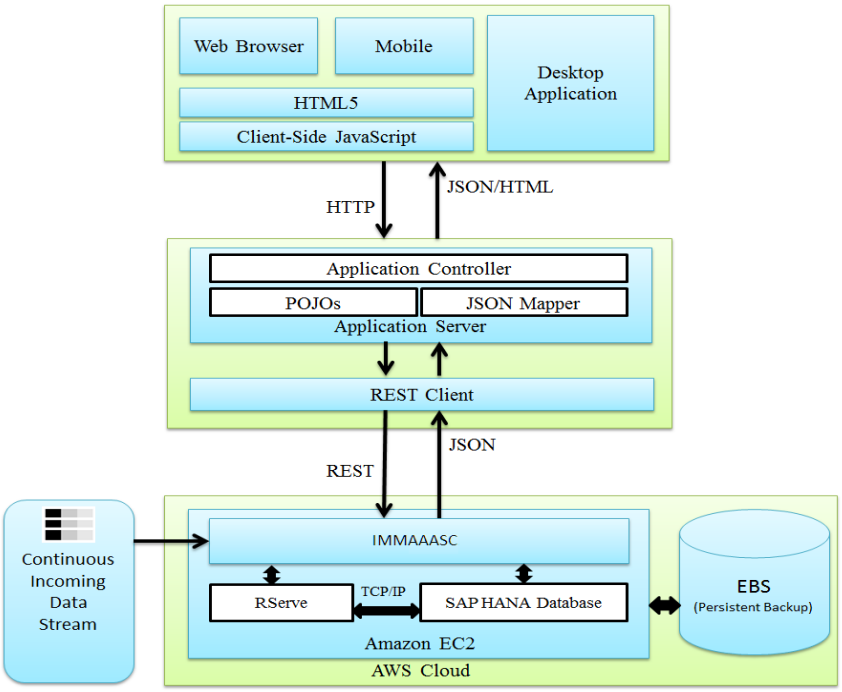

Figure 6. Real time analytics system architecture using IMMAAASC

VI. Conclusion

In-Memory Modeling and Analytics as-a Service in Cloud (IMMAAASC) platform is presented in this paper is an ongoing research topic where the usability aspect of an analyst is addressed by providing a sophisticated data manipulation tool in form of a service in cloud. Utilizing this platform we expect the analyst to be able to perform the business analysis tasks with minimal effort. The research is continued to make a stringent platform for fail-safe usage and secured transactions.

Acknoledgement

The authors would like to express gratitude for all the support and motivation provided by Mr. Ganapathy Subramanian, Vice President, TIP DNA (CESP), SAP Labs India Pvt. Ltd., Bangalore, Karnataka, India. We acknowledge his support for all the stimulating suggestions, encouragement and practical guidance during this research.

References

[1] C. Vecchiola, S. Pandey and R. Buyya, “High-Performance cloud computing: A view of scientific applications,” in *Proceedings 10th International Symposium on Pervasive Systems, Algorithms and Networks*, 2009, pp. 4-16.

[2] J. V öckler, G. Juve, E. Deelman, M. Rynge, and G. B. Berriman, “Experiences using cloud computing for a scientific workflow application,” in *Proc. ACM Workshop on Scientific Cloud Computing (ScienceCloud)*, 2011, pp. 15-24.

[3] X. Q. Yang and Y. J. Deng, "Exploration of cloud computing technologies for geographic information services," in *Proc. 18th International Conference on Geoinformatics*, 18-20 June 2010, pp. 1-5.

[4] C. Evangelinos, P. F. J. Lermusiax, J. Xu, P. J. Haley, Jr, and C. N. Hill, “Many task computing for real-time uncertainty prediction and data assimilation in the ocean,” *IEEE Transactions on Parallel and Distributed Systems*, vol. 22, issue 6, 2011, pp. 1012-1024.

[5] Cloud Foundry. (12 July 2013). [Online]. Available: http://www.cloudfoundry.com/.

[6] StarCluster. (12 July 2013). [Online]. Available: http://star.mit.edu/cluster/.
[7] RightScale. (12 July 2013). [Online]. Available: http://www.rightscale.com/.
[8] Heroku. (12 July 2013). [Online]. Available: http://www.heroku.com/.
[9] W. N. Venables, D. M. Smith, and the R Development Core Team, "An introduction to R," *Notes on R: A Programming Environment for Data Analysis and Graphics*, version 2.14.2, 2012.
[10] Amazon Elastic Compute Cloud (EC2). (2 July 2013). [Online]. Available: http://aws.amazon.com/ec2/.
[11] Amazon Elastic Block Store (EBS). (14 July 2013). [Online]. Available: http://aws. amazon.com/ebs/.
[12] Amazon Simple Storage Service (S3). (14 July 2013). [Online]. Available: http://aws. amazon.com/s3/.
[13] SAP HANA. (14 July 2013). [Online]. Available: http://www.sap.com/hana/hana-database/.
[14] SAP HANA One. (14 July 2013). [Online]. Available: http://www.saphana.com/docs /DOC-2437.
[15] Boto. (15 July 2013). [Online]. Available: https://github.com/boto/boto.
[16] Amazon EC2 Elastic IP Addresses. (14 July 2013). [Online]. Available: http://aws. amazon.com/articles/1346.
[17] Amazon Machine Images (AMI). (14 July 2013). [Online]. Available at: http://aws.amazon. com/amis.
[18] Bioconductor in the Cloud. (15 July 2013). [Online]. Available: http://www.bioconductor .org/help/bioconductor-cloud-ami/.
[19] R. Serve. (15 July 2013). [Online]. Available: http://www.rforge.net/Rserve/.
[20] RJDBC. (15 July 2013). [Online]. Available: http://cran.r-project.org/web/packages/ RJDBC/.
[21] S. Bhardwaj, L. Jain, and S. Jain, "Cloud computing: A study of infrastructure as a service (IaaS)", *International Journal of Engineering and Information Technology (IJEIT)*, vol. 2, issue 1, 2010, pp. 60-63.
[22] M. Boniface, B. Nasser, J. Papay, S. Phillips, A. Servin, K. Yang, Z. Zlatev, S. Gogouvitis, G. Katsaros, K. Konstanteli, G. Kousiouris, A. Menychtas, and D. Kyriazis, "Platform-as-a-Service architecture for realtime quality of service management in clouds," presented at the Fifth International Conference on Internet and Web Applications and Services, May 9-15, 2010, Barcelona, Spain.
[23] E. Knorr. (14 July 2013). Software as a Service: The Next Big Thing. [Online]. Available: http://www.infoworld.com/article/06/03/20/76103_12FEsaas_1.html.
[24] Amazon Web Services. (14 July 2013). [Online]. Available: http://aws.amazon.com/.

Prabal Mahanta was born in the city of Guwahati, Assam, India in 1985. Mr. Mahanta holds a Master's degree in Information Technology (M.Tech) with majors in Embedded Systems from International Institute of Information Technology (IIIT), Bangalore, Karnataka, India. He is currently working as a Developer at SAP Labs, Bangalore, India. He has worked extensively on critical customer projects as well as research projects. He has a strong inclination for research and teaching. He has presented many papers in national conferences and his research interest lies in in-memory computing, cloud-infrastructure, chaos theory and real time simulations.

Dhwanit Shah was born in city of Vadodara, Gujarat, India in 1989. Mr. Shah holds Master's degree (M. Tech) in Information Technology with majors in Database and Information Systems from International Institute of Information Technology (IIIT), Bangalore (Karnataka, India), as well as a Bachelor's degree (B.E.) in Computer Engineering from Vishwakarma Government Engineering College (VGEC), Gandhinagar (Gujarat, India) affiliated to Gujarat University (GU). He is currently working as a Research Fellow at SAP AG Innovation Center in Potsdam, Germany. He has previously worked as Software Developer at SAP Labs Bangalore, India for more than twelve months. Previously, he has also worked as an intern at SAP Labs for 6 months. He has research interests in database systems and software engineering.

