# Intelligent Big Data Analysis Architecture based on Automatic Service Composition

T.H.A.S. Siriweera, Incheon Paik, Banage T.G.S. Kumara, K.R.C. Koswatta
School of Computer Science and Engineering
University of Aizu, Aizu-Wakamatsu, Fukushima, Japan
siriweera@gmail.com , paikic@u-aizu.ac.jp , btgsk2000@gmail.com, ckoswatte@gmail.com

***Abstract*—Big Data contains massive information, which are generating from heterogeneous, autonomous sources with distributed and anonymous platforms. Since, it raises extreme challenge to organizations to store and process these data. Conventional pathway of store and process is happening as collection of manual steps and it is consuming various resources. An automated real-time and online analytical process is the most cognitive solution. Therefore it needs state of the art approach to overcome barriers and concerns currently facing by the Big Data industry. In this paper we proposed a novel architecture to automate data analytics process using Nested Automatic Service Composition (NASC) and CRoss Industry Standard Platform for Data Mining (CRISP-DM) as main based technologies of the solution. NASC is well defined scalable technology to automate multi-disciplined problems domains. Since CRISP-DM also a well-known data science process which can be used as innovative accumulator of multi-dimensional data sets. CRISP-DM will be mapped with Big Data analytical process and NASC will automate the CRISP-DM process in an intelligent and innovative way.**

***Keywords:*** **Big data analytics, Nested Automatic Service composition, Data mining, Architecture, CRSIP-DM**

## I. INTRODUCTION

Big Data is flooding by mass amount of data, which are generated by 2 types of sources those are man and unmanned. Data is pouring from every conceivable direction. In 2011, it was reported that the amount of information created and replicated nearly as many as bits of information in the digital universe is same as stars in the physical universe [1]. That means it shows exponential growth of digital data by various factors such as volume, velocity, variety, value, veracity etc.

In the same time importance and challenge to manipulate Big Data is increasing exponentially too. It has to be considered various factors and very difficult to synchronize all factors to make a final solid solution. Nowadays Big Data Analytics (BDA) is doing by manually accumulated tasks. Real-time analytical process is the most effective and efficient way of BDA.

BDA arises complex situation in its data mining process due to its diversified analytical requirements and multi-disciplinary data set's. We have to select a comprehensive data mining methodology to fulfil its diversified requirements in efficient way. We can choose a data-mining method for BDA and other necessary procedure using our experiences. CRoss Industry Standard Platform for Data Mining (CRISP-DM) is a useful standard for BDA. However, the manual process of steps in CRSIP-DM for BDA hinders faster decision making on real time application for efficient data analytics. Further, CRISP-DM has to pass thorough and rigorous steps to complete successfully the data-mining process.

In this paper, we propose a novel architecture to automate BDA process with CRISP-DM. We have identified two critical factors of the BDA, which are suitable technical framework and architecture. We have been selected Nested Automatic Service Computing (NASC) is key technology to automate CRISP-DM process. It has comprehensive capability to fulfil such multi-step process to automate while maintaining its scalability [2]. We assume Intelligent and Innovative integration of above-mentioned technologies will result a scalable intelligent real-time BDA solution.

The rest of this paper is organized as follows. In Section II we discuss about preliminaries, in Section III we discussed about architectural designs, in Section IV we discuss about the evaluation, in Section V we discuss related works. Finally, in Section VI we conclude the paper.

## II. PRILIMINARIES

### A. Big Data Analytics Process

BDA is the process of collecting, organizing and analysing large sets of data to discover patterns and other useful information. BDA helps organizations to better understand the information contained within the data and help to identify the data that is most important to the business and future business decisions. Data warehouse will be processed by data science technology and mined by data mining techniques. Data manipulation will be done by a data science process and here we will use, CRISP-DM as shown in Fig. 1.

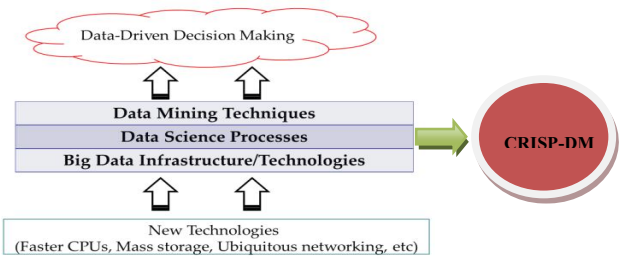


**Figure 1.** Standard Process of BDA

978-1-4673-7278-7/15 $31.00 © 2015 IEEE
DOI 10.1109/BigDataCongress.2015.46

### B. CRISP-DM Process

CRISP-DM has six stages and the stages fit to address effectively to complete data science requirement's of the Big Data domain. Fig. 2 shows the graphical view of the model.

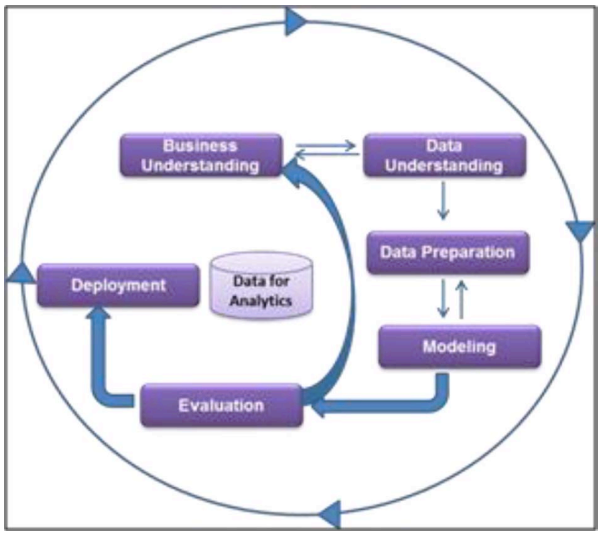


**Figure 2:** CRISP-DM Process

In **Business Understanding stage**, we understand the objectives and requirements from a domain perspective and a preliminary plan and designed to achieve the objectives.
**Data Understanding stage** initiates with given data set and continue with tasks up-to discover first insights into the data. **Preparation stage** prepares final purified and rectified data set needs to be moved to next stage. **Modeling stage** will apply various modeling techniques usually data mining techniques based on the requirements. **Evaluation stage** does thorough insight about the model with matured data. End of this stage it can take a decision whether it is use the mining process results. In **Deployment stage,** the result will be organized to present according to the customer readable and deploy the project.

We note Business understanding and data understanding are manually confirmed and already done by this project. Next we have to deal with rest of four stages to automate the process by NASC technology.

### C. Nested Automatic Service Composition

NASC [2] is a technology derived from service oriented architectural design pattern. In this research, we use NASC to automate BDA. To automate BDA process in intelligent way, we need to define concepts of services for each step of the CRISP-DM process. After, each step will be matched to composition step logically. Development of intelligent BDA process includes these steps:

(1) Service types and instances development for the BDA; (2) Define workflow for BDA; (3) Service discovery algorithm is developed for BDA; (4) Service selection algorithm is developed for BDA and (5) service algorithm is developed for BDA result.

## III. ARCHITECTURE FOR INTELLIGENT BDA

Our problem domain is to make a comprehensive architectural solution. That is interpreting to real world problem in to the technical language. Big Data solution engineering perspective (size and concerns) and due to complexity of the solution, critical time-to-market needs demand new software engineering approaches to design software architectures [3]. One of these approaches is software Reference Architecture (RA) that eases to systematically reuse's knowledge and components when developing a concrete System Architecture (SA) [4]. Finally we have been able generate implementation level UML class diagram easily.

### A. *Reference Architecture*

RA is an architectural solution, which facilitate to make template solution for complex problem domain. RA for the BDA process is shown in Fig. 3. It provides solid base to extract SA from that. SA is a conceptual model that defines structure, behaviour and more views of a system. Simply RA is layered solution, which gives high-level view how each components and technologies of the product behave and how it maintains interactions between each of them. This layered pattern connected closely to an architectural principle "loose coupling" [5].
For RA perspective, we have identified main 3 building block layers, top level layer called as *Analytical Layer*, middle layer called as *Technology Layer* and bottom layer called as *Infrastructure Layer*. Let start to identify each layers in summarily:

**Infrastructure layer:** It mainly considers data warehouse and data mart layer. This contains Hadoop Eco system to manipulate Big Data infrastructure, web service pools and two relational data base managements (RDBMS) for data manipulation and to maintain analytical clusters. All of the above can be existed in Intranet and Internet platforms. As an example, Hadoop cluster can be Geo distributed as data centres and then we have to deal with hadoop beyond the intranet level.

Since web services also can be distributed along the internet and local network. One of two RDBMS is to accepted export data from Hadoop and data processing facility to the technology layer. The other RDMS is used for handling analytical cluster and related activities of the analytical process.

**Technology layer:** Mainly this layer dominated by NASC and it supports technologies such as quality of services agent and intelligent planning agent to provide intelligent workflow automation facility.

Therefore it will identify the requirement utilize respective resources distributed along the system to fulfil the functional and non-functional requirements of the projects.

**Analytical layer:** This layer dominated by CRISP-DM to provide data mining process of the project. First two out of 6 stages of the CRISP-DM has been decided by manually and therefore ASC will be dealt with only the rest of four stages

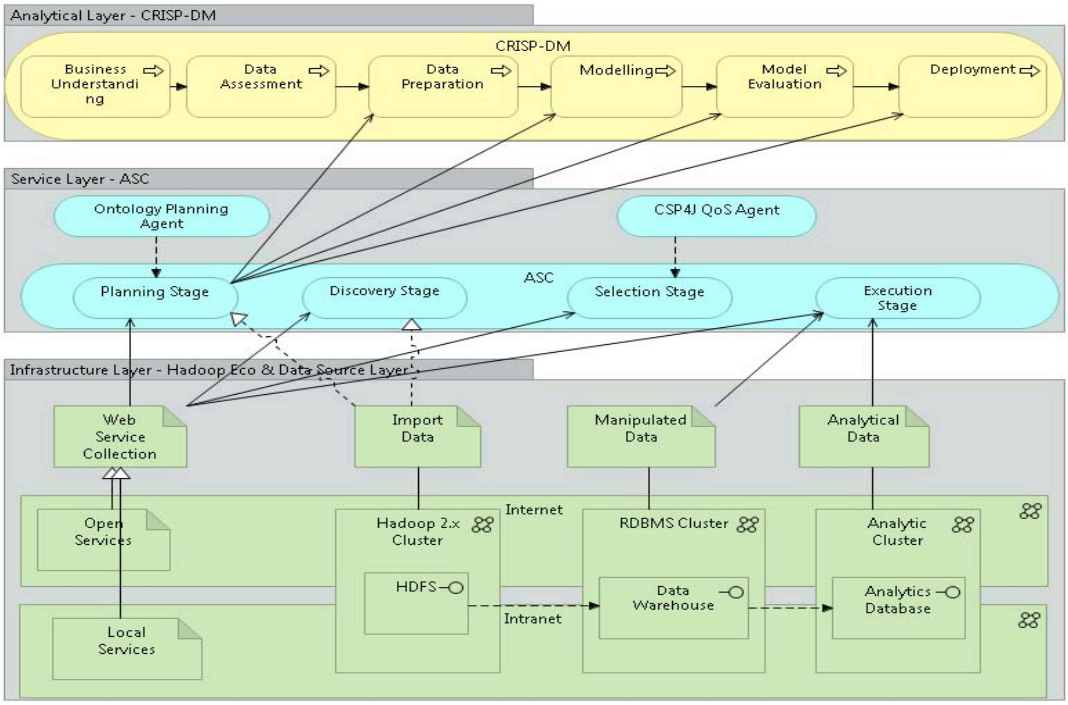


**Figure 3.** Reference Architecture of the BDA Solution

to full-fill the data automation of mining requirement.

### *B. System Architecture*

**Scenario 1:** *ABC Air Port Company requires analysis of the flight delay data to identify factors, which had been affecting to the flight delay. By the analysis, company hopes to take necessary decisions to reduce/ avoid flight delays.*

We derive SA based on the RA and applied our scenario to the SA. It is clearly shown how each layer will behave during the execution time and result (output) will be produced in it is (NASC) execution stage.

As shown in Fig. 4, it is giving high-level view how we have derived SA based RA applied with our scenario. Below SA is displaying the existing technologies and their responsibilities. Moreover, it describes communication between layers across the total solution.

### *C. Top Level UML Class Diagram*

We are able to successfully integrate main technologies to achieve intelligent real-time analytics for BDA by using RA. Next, we successfully derived SA for our scenario from the RA. Finally we could be able to extract and designed detailed top level UML class diagram of scalable BDA for our scenario based on the RA and the SA.

Mainly we have identified 2 packages. One is for ASC, the other is for CRISP-DM related services.

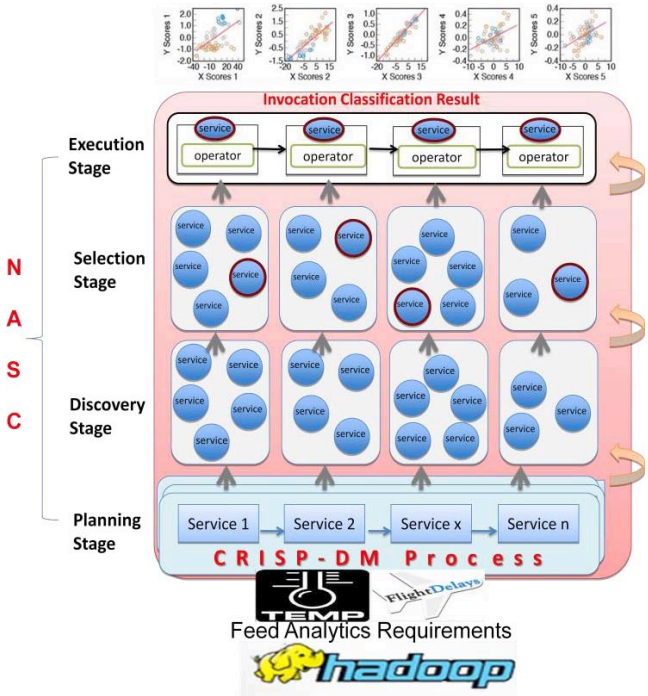


**Figure 4.** NASC vs CRISP-DM and Classification Results

In addition to them, there are two utility packages to provide utility services to system. They are Planning Agent and Quality of Service (QoS) Agent.
Figure 6 displays high level view of the UML class diagram.

**NASC Package**
This is the based package of the solution. This allows identifying the functional and non-functional requirements for the analytics as follows.

R: Set of user's requests at the service level.
$W = \{t_1, t_2, t_3, \ldots t_l\}$: Set of $l$ abstract tasks in an abstract workflow W.

Planning: $\prod$: R → W.
$I_i = \{i_{i1}, i_{i2}, i_{i3}, \ldots, i_{im}\}$ : Set of $m$ service instances advertised in a service registry for an abstract task $t_i$. I is the set of $I_i$, for $1 \leq i \leq l$. If each task in a workflow has $m$ instances, then the total number of service instances available for the workflow is $l \times m$.

Discovery: $\Delta : W \rightarrow I$
$C_j = \{c_{j1}, c_{j2}, c_{j3}, \ldots, c_{jp}\}$ : Set of $p$ selected service instances to be executed from the service instance set $I_i$. C is the set of $C_j$ where $1 \leq j \leq l$

Selection: $\Sigma : I \rightarrow C$.
$X = \{x_1, x_2, x_3, \ldots, x_q\}$: Set of $q$ executed service traces.

Execution: $E : C \rightarrow X$

**CRISP-DM Package:** This package will be responsible to deal with web services (internet and locally distributed) related to services which are requesting by NASC to accomplish complex, dynamic and diversified tasks of the BDA process.
Note that according to the scenario, we have manually accomplished 1st 2 stages of CRISP-DM, here NASC will deal with the rest four stages to be automated.

**Utility Packages**: One of the two main packages is planning agent. It can be selected developer preferred planning agent to full-fill the planning requirement such as HTN. However, we choose Planning Agent by Ontology reasoning for the planning process.
Quality of Services (QoS) Agent: We have used Constraint Satisfaction Problem Solving Agent as our QoS agent.

## IV. EVALUATION

We observed and studied the following as main advantages,

- NASC is a technology to create scalable solution for real-time analytics solution. The approach to use ASC is useful to automate CRISP-DM process, because it separates workflow management from functional modules. We assume this will be technically more effective solution than conventional manual path.
- Technologies used in RA can customize according to the user's preferences.

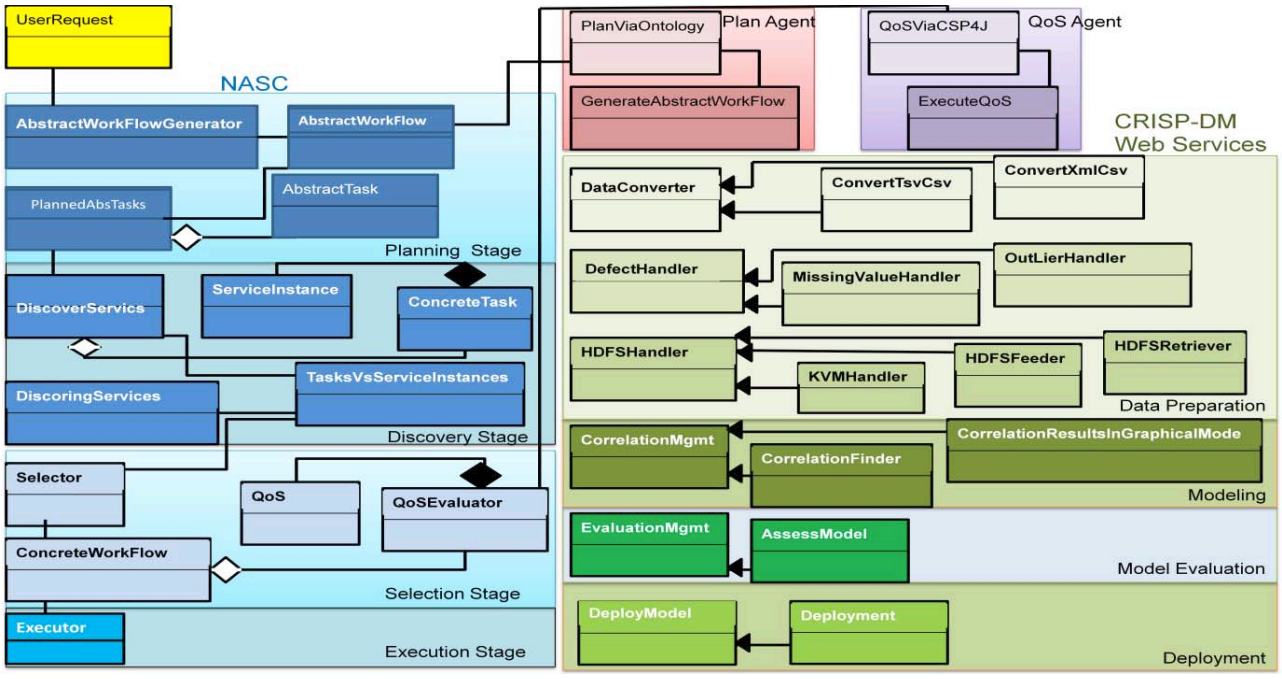


**Figure 5.** Top-level UML class diagram of Intelligent BDA Process

- This is layered architectural solution. Therefore, from architectural perspective solution is loosely couple and interoperable.

## V. RELATED WORK

The literature on scalable intelligent architecture for BDA is scarce. There is an architecture, which provided reference architecture for BDA, and result is indicative evidence [6]. In addition, intelligent multi agent solution provided for particular domain [7].

A memory centric real-time BDA also introduced and explained [8]. Health related real-time BDA solution for monitoring purposes discussed in detail [9].

And it is providing predictive BDA for aviation industry with considering various factors of the aviation industry [10]. another research team has introduced a method to make informed predictions, and gain business insight from the steady influx of information within various business domains [11].

Wu et al. [12] presented a HACE theorem that characterizes the features of the Big Data revolution, and proposes a Big Data processing model.

Most of solutions are domain specific and some of them are providing real time support to the analytical process. When it is comparing with our solution, we have introduced a domain independent scalable solution to the BDA process.

## VI. CONCLUSION

We have achieved successful design for Intelligent BDA using RA. Next, we derive the SA using that RA and simulate SA with our scenario. After that, we design UML class diagram for the software development process of BDA. As the solution is scalable and we believe this architectural solution will work for our scenario effectively.

In our future work, we need to do more experiments to evaluate about efficiency.

## REFERENCES


[1] IDC Digital Universe Study, sponsored by EMC, June 2011,http://chucksblog.emc.com/chucks_blog/2011/06/2011-idc-digital-universe-study-big-data-is-here-now-what.html

[2] I. Paik, W. Chen, and M. N. Huhns, "A scalable architecture for automatic service composition," IEEE Transactions on Services Computing, vol. 7, no. 1, pp. 82–95, Jan.–Mar. 2014.

[3] Nakagawa, E., Olivera, P., Becker, M.: Reference architecture and product line architecture: A subtle but critical differences. ECSA, pp. 207-211 (2011)

[4] Cloutier R., Muller G., Verma D., Nichiani R., Hole, E., Bone The Concept of reference architectures, System Engineering 13(1), 14-27 (2010)

[5] P. Avgeriou en U. Zdun, Architectural Patterns Revisited - A Pattern Language," 10th European Conference on Pattern Languages of Programs, Irsee, Germany 2005

[6] B. Geerdink, I.G roep – "A Reference Architecture for Big Data Solutions" – IEEE ICITST-2013

[7] M. Ivan, M. Stula, J Maras – Intelligent MultiAgent System for Insurance Industry – IEEE 2014 MIPRO, Crotia

[8] Z. Tao, K. Doshi, X. Tang, T.Lou, Z. Lu, H. Li – "A Big Data Architecture for Real-time Analytics"– IEEE 2013 Big Data

[9] P. Moore, L. Baroli, F. Xhafa, A.Thomas – "Monitoring and Detection of Agitation in Dementia. Towards Real- Time and Big-Data Solutions" – IEEE 2013 Computer Society

[10] S. Ayhan, J. Pesce, P. Comitz, D. Sweet, S. Bliesner, G. Gerberick – "Predictive Analytics with Aviation Big Data" – IEEE 2013 ICNS Conference

[11] http://www.oracle.com/technetwork/database/options/advanced-analytics/bigdataanalyticswpoaa1930891.pdf:Access2015.02.13

[12] X. Wu, X. Zhu, G. Q. Wu, and W. Ding , "Data Mining with Big Data," IEEE Transaction Knowledge and Data Engineering, vol. 26, no. 1, pp. 97–107, 2013.