International Journal of Research
Available at
https://edupediapublications.org/journals

p-ISSN: 2348-6848
e-ISSN: 2348-795X
Volume 04 Issue 05
April 2017

A Rapid Quality-Aware Development of Data-Intensive Cloud Applications

Divya Byri
Department of CSE

ABSTRACT: *In this paper, we deliberate the query of howquality-aware MDE should provision data-intensive softwaresystems. This is a difficult challenge, since current models andQA techniques largely ignore properties of data such asvolumes, velocities, or data location. Additionally, QA necessitatesthe ability to characterize the behavior of technologies such asHadoop/MapReduce, NoSQL, and stream-based processing,which are poorly understood from a modeling standpoint. Tofoster a community response to these challenges, we presentthe research agenda of DICE, a quality-aware MDEmethodology for data-intensive cloud applications. DICE aimsat developing a quality engineering tool chain offeringsimulation, verification, and architectural optimization for BigData applications.*

KEYWORDS-Big Data, quality assurance, model-drivenengineering

I. INTRODUCTION

Massive popularity and wide-scale deployment ofthe Internet has enormously increased the rate ofdata generation and computation [5, 6]. Thishuge growth has also highlighted immense potential for utilization and analysis of data overa wide set of users and its applications. Consequently, unprecedented data-related challengeshave emerged.Consider an example of a simple Internetsearch engine that ranks documents on the basisof relative frequency of search terms in its datacollection. The search engine could be enhancedif it includes consideration of user-clicks whileobtaining popular results. Similarly, geographicallocation of users could be incorporated to increaserelevancy. The two enhancements mentioned heremay seem plausible; however, considering themassive dataset of Internet documents and the diverse geo-location of Internet users, they requirecomprehensive collection, efficient storage andretrieval, extensive linkage, meticulous investigation, and methodical analysis; most importantly, ina precise and timely manner. Further, extensiveequirements of meeting availability, scalability,and high performance also exist.The extensive challenges mentioned aboveare not restricted to search engines. With theemergence of clouds, the notion of computinghas incorporated new requirements of providingefficient user access and storage [80]. Further, theterms of availability and scalability are inherentwith cloud systems. In addition, for a multi-usersystem, a cloud system needs to fulfill the requirements of privacy and access controls.In the data-intensive world we live, requirements and challenges also vary with applications,For example, an iterative application such aspage-rank computation algorithm requires iterative computation until a point of convergence isreached. In comparison, a streaming applicationwould prefer processing stream of events in orderto provide timely results.

II. RELATED WORKS

Data Intensive computing refers to computing oflarge scale data. Gorton et al. describe types ofapplications and research issues for data intensivesystems. Such systems may either includepure data-intensive systems or they may also contain data/compute-intensive systems. In that, theformer type of systems devote most of their timeto data manipulation or data I/O, whereas in thelatter type data computation is dominant. Normally, parallelization techniques and high performance computing are adopted to encounterthe challenges related to data/compute-intensivesystems.With the growth of data-intensive computing,traditional differences between data/computeintensive systems and pure data-intensive systemshave started to merge and both are collectivelyreferred as data-intensive systems. Major researchissues for data-intensive systems include management, handling, fusion, and analysis of data. Often, time-sensitive applications are also deployedon data-intensive systems.

Depending upon its usage, a data-intensivecloud could either be deployed as a privatecloudsupporting users of a specific organization, or it may be deployed as a public

International Journal of Research
Available at https://edupediapublications.org/journals
p-ISSN: 2348-6848
e-ISSN: 2348-795X
Volume 04 Issue 05
April 2017

cloudproviding shared resources to a number of users.A data-intensive cloud entails many challengesand issues. These include data-centric issues suchas implementing efficient algorithms and techniques to store, manage, retrieve, and analyzethe data and communication-centric issues such asdissipation of information, placement of replicas,data locality, and retrieval of data. Note that issuesin the two categories may be interrelated. For instance, data locality often leads to faster executionof data.

Grossman and Gu [7] discussed varieties ofcloud infrastructures for data intensive computing. Fig.1 illustrates the two architectural models for such a system: a cloud could provide EC2-like instances for data-intensive computing, or itcould offer computing platforms (like MapReduce) to its users. In the former case, a user isrequired to select tools and a platform for computing, and the cloud provider is responsible forstorage and computing power. The provider is alsoliable for replication, fault tolerance, and consistency. In comparison, for platform-based cloudcomputing, application-specific solutionsexist which provide enhanced performance.

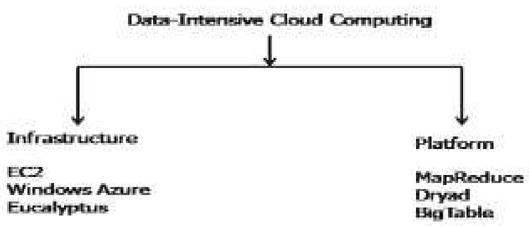

Fig. 1 Architecture model of data-intensive cloudcomputing

In this paper, we mainly resort to the lattercategory (data-intensive computing platforms) asthey specifically address challenges and solutionsto data intensive computing. However, duringthe paper, we discuss a few infrastructure-relatedissues such as effective network utilization andresource sharing which may well be applied toboth the types.

III. APPROACHES

The corearea of DICE is to define an MDE approach and aQA tool chain to continuously enhance data-intensive cloudapplications with the goal of optimizing their service level, we believe thatthe methods and tools shown in Table 1 are required toprovide a compresensive quality-aware MDE approach forBig Data applications. The DICE IDE will guide thedeveloper throughout this methodology. From these models, thetool chain will guide the developer through the differentphases of quality analysis (e.g., simulation and formalverification), deployment, testing, and acquisition offeedback data through monitoring. This data will then beprocessed and fed back to the IDE through the iterativequality enhancement tool chain, which will analyze runtimedata to detect quality incidents and anti-patterns in theapplication design. This will provide feedbacks to guide thedeveloper through cycles of iterative quality enhancement.

A. DICE Profile: MDE for Data-Intensive Applications

Models in DICE should be formulated at three levels, calledDPIM, DTSM, DDSM, which we deliberatesubsequent.

DICE Platform Independent Model (DPIM). The DPIMmodel corresponds to the OMG MDA PIM layer anddescribes the behavior of the application as a directedacyclic graph that expresses the dependencies betweencomputations and data. This model should also expresssource data formats, synchronization mechanisms in thecomputation logic, and quality requirements for bothcomputation logic and data transfers.

Fig.2 shows a possible example of DPIM for anapplication including four Data Sources (DS1-DS4) andfour Computational Logic elements (CL1-CL4). At theDPIM layer the designer can specify the data format (e.g.,structured or semi-structured data, flat files, etc.) andindicate if the data is transferred between processing stepsvia a shared storage system (e.g., S1) or obtained from datastreams (e.g., DS3 and DS4 flows).

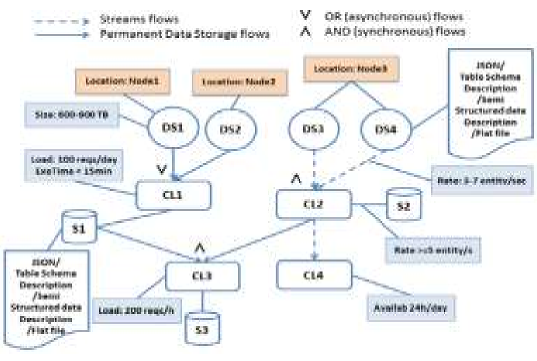

Figure 2. DICE platform independent model (DPIM)

International Journal of Research
Available at
https://edupediapublications.org/journals
p-ISSN: 2348-6848
e-ISSN: 2348-795X
Volume 04 Issue 05
April 2017

A computational logicelement can process multiple flows both synchronously orasynchronously. Data locations, estimated size (e.g., 600-900 TB for DS1), computation logic workload (e.g., 200requests/h for CL3) and service-level constraints (e.g., CL1runtime less than 15 minutes) may also be specified.

DICE Platform and Technology Specific Model (DTSM). ADTSM, illustrated in Figure 3, consists of a refinement ofthe DPIM and includes some technology specific concepts,both for computational logic and data storage, but that arestill independent of the deployment. For example, data andcomputational logic elements may be associated at theDTSM layer with specific technologies. DS1 and S1 maybe required to be based on the Hadoop File System (HDFS),DS2 on a relational database (RDBMS), CL2 on complexevent processing (CEP), and so forth.

Table 1. DICE Tools

DICE profile	A novel data-aware UML profile to develop data-intensive cloud applications and annotate the design models with quality requirements.
DICE IDE	Integrated development environment with code generation to accelerate development.
Quality analysis	A tool chain to support quality-related decision making composed by simulation, verification and optimization tools.
Iterative quality enhancement	A set of tools and methods for iterative design refinement through feedback analysis of monitoring data.
Deployment and testing	A set of tools to accelerate deployment and testing of data-intensive applications on private and public clouds.

DICE Platform, Technology and Deployment SpecificModel (DDSM). The DDSM, shown in Figure 4, is aspecialization of the DTSM model which adds informationabout the technology in use and the application deploymentcharacteristics. For example, the deployment may bespecified at the DDSM layer with details on the systemcapacity (e.g., CL1 will be hosted on 50 EC2 ElasticMapReduce xlarge instances).

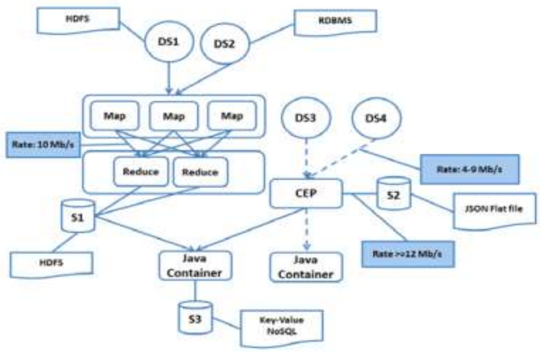

Figure 3 DICE Platform and Technology Specific Model (DTSM)

DICE will help the developerdeciding deployment characteristics by identifying throughnumerical optimization a deployment plan of minimum cost,subject to performance and reliability requirements.Additionally, deployment tools will be able to process theinformation provided by the DDSM to minimize the effortrequired to deploy the application. Transformations betweenDPIM, DTSM and DDSM models will be supported by theDICE tool chain.

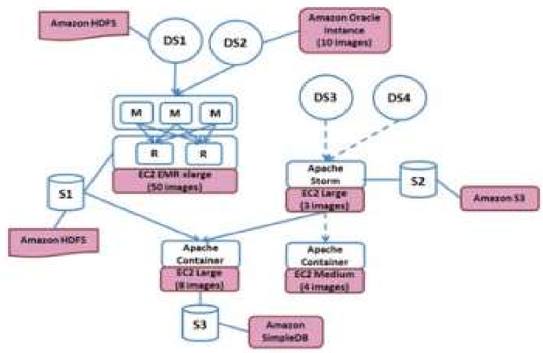

Figure 4 DICE Platform, Technology and Deployment SpecificModel (DDSM)

B. Quality Annotations

The DICE profile will enable the design of data-intensivecloud applications. In particular, as highlighted in Section II,we envision that the DICE profile needs to include at least:

(i) quantitative annotations on the availability of a datasource or intermediate by-products resulting from a datatransformation;

(ii) annotations to specify rates, latenciesand utilizations of resources, including the possibility to specify service level constraints on data transfers;

(iii)annotations to specify costs of data-intensive applications;

(iv) safety annotations that will be treated as hard constraints.

C. Deployment

The last set of requirements for the DICE approach to beeffective concerns the development of appropriate tools tosupport the application deployment and initial testing.Ideally, the primary target of an MDE methodology for BigData should be either private cloud applications or publiccloud applications that can use cloud platform services forBig Data, such as Amazon Elastic MapReduce or cloudbased storage services. Automatic deployment andconfiguration from DDSM models could be achieved usingextensions of tools such as Brooklyn, Puppet or Chef.

IV. CONCLUSION

We have designated the investigationprogram of DICE, a visionfor a novel model-driven engineering approach preciselytailored to Big Data applications. We have recognized severalchallenges that arise in this area due to limitations in currentmodels and quality analysis tools that arise from theinability to fully describe data operations and datacharacteristics.

REFERENCES

[1]. Abadi, D.: Data management in the Cloud: limitationsand opportunities. In: IEEE Data Engineering (2009)

[2]. Abadi, D.: Problems with CAP and Yahoo's littleknown NOSQL System. Available.http://dbmsmusings.blogspot.com/2010/04/problems-with-cap-andyahoos-little.html. Last accessed 4 Oct 2012

[3]. Abe, Y., Gibson, G.: pWalrus: Towards better integration of parallel file systems into cloud storage. In:Workshop on Interfaces and Abstractions for Scientific Data Storage (IASDS10), co-located with IEEEInt. Conference on Cluster Computing 2010 (Cluster10), Heraklion, Greece (2010)

[4]. Abouzeid, A., Bajda-Pawlikowskim, K., Abadi, D.,Silberschatzm, A., Rasin, A.: HadoopDB: An architectural hybrid of MapReduce and DBMS technologies for analytical workloads. In: VLDB (2009)

[5]. Gorton, I., Greenfield, P. Szalay, A., Williams, R.:Data-intensive computing in the 21st century. IEEEComputer 41(4), 30–32 (2008)

[6]. Kouzes R., Anderson G., Elbert S., Gorton, I., Gracio,D.: The changing paradigm of data-intensive computing. IEEE Computer 42(1), 26–34 (2009)

[7] D Ardagna, E Di Nitto, et al. MODAClouds: A modeldriven approach for the design and execution of applications onmultiple Clouds, Proceedings of MiSE 2012, 50-56.

[8] S. Bernardi, J. Merseguer, D. C. Petriu. Dependabilitymodeling and analysis of software systems specified with UML.ACM Computing Surveys, 45(1), p. 2, 2012.

[9] P. Debois. Devops: A software revolution in the making?,J. Information Technology Management, 2011

[10] D. A. Menascé, J. M. Ewing, H. Gomaa, S. Malek, J. P.Sousa. A framework for utility-based service oriented design inSASSY. Proceedings of ACM/SPEC WOSP/SIPEW 2010, 27-36.

[11] A. Martens, H. Koziolek, S. Becker, R. Reussner.Automatically improve software architecture models forperformance, reliability, and cost using evolutionary algorithms.Proceedings of ACM/SPEC WOSP/SIPEW 2010, 105-116

[12] D. Franceschelli, D. Ardagna, M. Ciavotta, E. Di Nitto.Space4Cloud: A tool for system performance and cost evaluation ofcloud systems. Proceedings of MultiCloud workshop, 27-34, 2013.

[13] J. F. Perez and G. Casale.Assessing SLA compliance fromPalladio component models.Proceedings of the 2nd Workshop onManagement of resources and services in Cloud and Sky computing(MICAS), IEEE Press, 2013.