Development Planning in the Big Data Era: Design References Architecture

Wael ALzyadat, Aysh Alhroob

ABSTRACT--- Big data concept which scale for amount data, that generated from several data sources through captured data process produced large dataset from multiple domains; cloud platforms provide the scalability and availability to carrying on volume, management and analytics data. Software concept in big data regard to cloud deployment involve three main layers are data layers (sources data), data aggregation, and analytics layer. In this research we proposed development references architecture which indicates the processes through life data cycle, cloud (SaaS,PaaS, and IaaS), and analytics layer. The preprocess and acquisition process drive interlinking among layers and provide the validate process.

Keywords—Big Data, Design, Reference Architecture, Cloud, Preprocess, Transformation

1.0 INTRODUCTION

In recent years, big data is a strong impact in every sectors and industry. Especially, interdisciplinary disciplines are considering to big data era, with rise of digital universe rapidly expanding by created, captured, sharing, broadcasting and replicated. In 2000 Wal-Mart produced 110 Terabytes, by 2004 recorded 500 petabytes, until 2005 the storage capacity was traditionally treating with information. Forward to 2006, amount of digital information was 161 Exabyte that amount close with storage scale; by 2007 it surpasses about 9 Exabyte of storage capacity ,in 2010 by 400 Exabyte [1]. IDC report [2], commissioned by Dell EMC, predicts more than 40 zettabytes of digital data in 2020. The period of 2001 to 2008 was the evolutionary stage for big data development, that encourage academics and practitioners be interested in understanding about the impacts of big data terms. Example, search rose topic by google search engine was 252000 hits on November 2011, five months later was 1.39 billion hits on April 2012 which increased about1.389748 hits; after twenty months' period the rose topic almost 1.69 billion hits on December 2013; the amount of hits is from November 2011 to April 2012 increased about 1.389748 billion hits as well as the amount increased from April 2012 to December 2013 about 0.3 billion [3].

Another example is YouTube received 28 hours of video every 60 second on 2011 [4], at 2018 received 300 hours of video every 60 seconds which increased about 272 hours [5]. Due to examples are led to wave of big data via quantity of storage. Obviously, the *Big* term are pretend from huge amount of data. Which references to characterize of big data in volume.

Volume is a backbone for big data concept which scale for amount data, that generated from several data sources through captured data process produced too large dataset from multiple domains such as social media, devices, historical data, and archived data; to elucidating term of volume is dealing with large scales of data within data processing [6].

Data warehouse organizes the data in the repository by collection from several databases but should be pre-determined. On other hand, the weak request organizes streaming data such as mobile devices and click streaming, the big data promise to treat with weak from data warehouse.

Dive to volume from big data concept as aforementioned example for rose topic, in short period time (5 months) the hits amount increased 5515.8 times compare from 252000 hits on November 2011. While, after twenty months increased to 1.389748 hits on April 2012. Meanwhile, 1 billion users used google products every month, Obviously, the infrastructure such as datacenter network equipment's and cooling system. Which lead to boosting 14 datacenters across the US, Europe, and Asia linked with billions of distributed devices [7].

The effective of volume from back end stage, required to upgrade datacenter via different aspects storage capacity, telecommunication, operation and deployment; to capable datacenter be harmony and uptake with data era should focus on modernization components such as Internet Data Center (IDC) and Cloud Data Centre (CDC). Consequence, the front-end stage of volume term is many data sources congregation that producing at the repositories at the data center by differ vastly from those at the storage capacity.

The data sources are maestro on which data paradigms and mechanism need to use it, the primary data that captured or collected from source named raw data; a sensor device purpose is to detect event or change in environments ingests data over time, the sensors classified based on memory storage such as GPS provider with data logger.

That impetus IT industry competitive into potential market such as Microsoft, Google, and Oracle, and other IT companies [8].

Microsoft produced Azure Data Lake [9] capabilities to work with any size of data (volume), analysis, management, and running parallel (velocity) and transformation via batch and/or streaming data (veracity). Google Cloud Platform (GCP) [10] distributed dataservices carry on big data perspective, by integration into GCP which the volume (amount of data) use google cloud storage, the management structure, semi-structure and/or unstructured provide by BIGQUERY and stackdriver.

Revised Manuscript Received on May15, 2019.
WaelALzyadat,India
AyshAlhroob, India

Oracle Cloud Service [11] provide block storage volumes by two components are instance, volume attachments via network communication (iSCSI) and virtualization (Paravirtualized).

The cloud platforms provide the scalability and availability to carrying on volume dimensional. Meanwhile, the management and analytics data.

Gartner report titled Magic Quadrant for Data Management Solutions for Analytics [12], indicate the scalability of volume , extract, transforming, and loading.

Figure 1 Data Management Solutions for Analytics (DMSA)
(Source: Gartner, 2019)

References architecture significance to understand typical functional (method) in abstraction level and the data flow (management) level as manipulate data, transform, and loading [13], the big picture of it how to practicality based on big data analysis.

For cloud computing references architecture consist of services deployment, services orchestration, cloud services management, security, and privacy [14].

The services oriented of references architecture for cloud has been presented by analysis volume data as shown in figure .1 through scale ***X***: completeness of vision and ***Y***: ability to execute.

The big data system relies on cloud platform to achieve the dimensions are 4Vs, almost of cloud paradigms use integration product/services through three major layers as provider are: 1) Infrastructure as a Service (IaaS), which cover the volume (storage capacity) from big data dimension such as Oracle implement by iSCSI and Paravirtualized. 2) Platform as a Service (PaaS), cover the velocity and management data to deploy use network and communication, such as Google provide BIGQUERY and stackdriver, and 3) Software as a Service (SaaS), provide the consumer accessibility running into IaaS by interfaces or API. Due to, configure throughput PaaS played as middleware, Such the Microsoft product name Azure Data Lake from both batch and streaming; the capability to assigns for veracity and velocity.

The goal of this research are: 1) planning design for references architecture big data in scope of cloud computing SaaS, PaaS, and IaaS. 2) Identifies the major's activities and dimensions for big data demand on structure and model for acquisition information.

2.0 REFERENCE ARCHITECTURE FOR BIG DATA

Section 2.1.present earlier taxonomy of big data term. Research on latest report of planning of design reference architecture are presents in section 2.2. Finally, the Structure and Model for big data section present in 2.3.

2.1Taxonomy of Big Data

Big data analytics architecture is rooted in the concept of data life cycle framework: data capture, proceeds via data transformation, and data Consumption; presents in three layers. First layer, data Layer consist three major kinds structured data (traditional electronic), semi-Structured (Log of health monitoring device), unstructured (Clinical Image). By two faces internal and external which stored depending on the content format[15] [16].

Second layer is data aggregation to handling various data source throughout, A) Acquisition is read data from various communication channel frequencies, size, and format; B) Transformation engine, cleaning, monitoring, splitting, translating, merging, sorting, and validating. C) Storage.

Third layer is analytics, capability in the context to acquire, store, process, and analyze large amount of data in various forms, and deliver meaningful information to user that allow them to discover values and insight in a timely fashion [17].

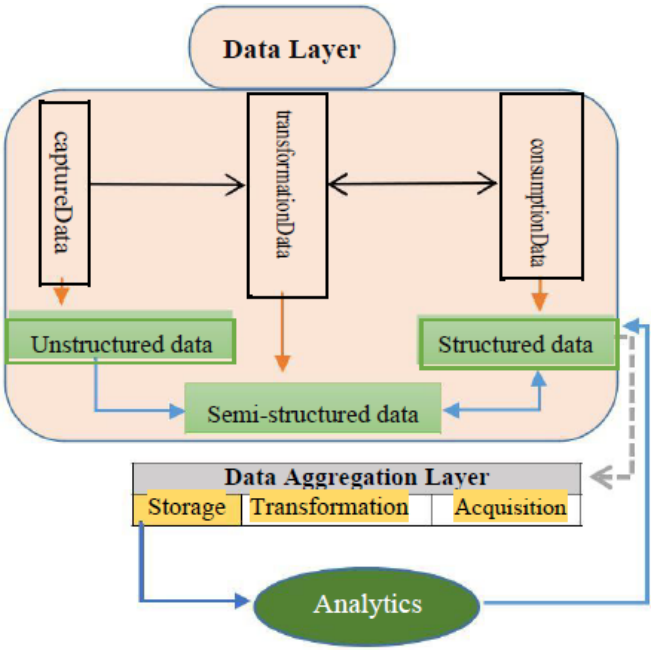

Figure 2 Data life cycle framework

Analytics categorize aspirational, experienced, and transformed. While, the act of it such predictive analytics and traceability are request data consistent, visible and easily accessible for analysis (suitable and scalability) Potential benefits of big data analytics IT infrastructure, operational benefits, organizational, managerial, and strategic.

2.2Reference Architecture

Big data architecture used in different waves, one of this wave is smart grids which appear the significance to focus on data model.

The firstresearch attempt to apply big data into smart grid was in 2015 [18] titled a big data architecture design for smart grids based on random matrix theory. Subsequent, power system generates raw data under big data umbrella. Divided by three layers as show figure.

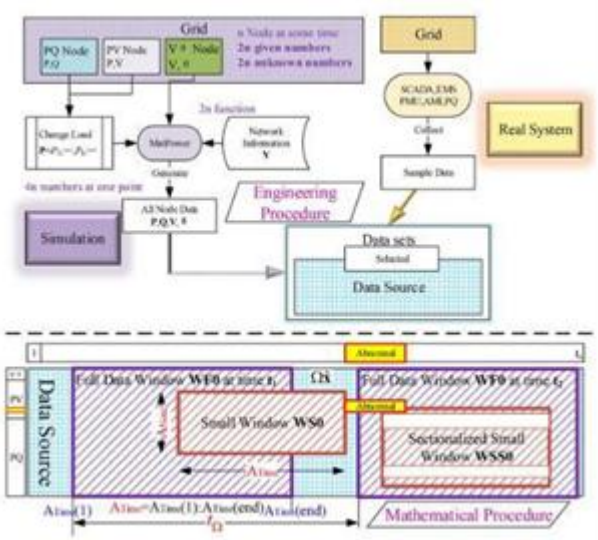

Figure 3 Designed big data architecture for smart grids, source [18]

The first layer is big data architecture for smart grids, use mathematical procedure for data source; the second layer is advantages in data process use interrelation and interaction analyses via mathematical procedural. Final layer is advantages in management mode throughput data flow implement by decentralized control system, cloud computing.

2.3Structure and Model for Big Data

The DICE Methodology, adopt the modeling standard for data-intensive applications [19], used four main components are : DPIM pinpoint to volume by two nodes storage node scale of capacity and compute node for input data; the second component select the compute node to Cassandra cluster by apply refinement concept; to prepare schema between two nodes internally using Hadoop (MapReduce) that achieve in DTSM component. The DDSM component physical mapped nodes in TOSCA blueprint via cloud application.

Model Driven Engineering (MDE) techniques are frequently applied for big data analytics MDE: Visual representation, documenting and sharing idea; which reducing the level of abstraction allow to transformed to low level targets model; that lead to perform analysis and verification tasks in early stage. the challenge for the scientists and other predictive professionals due to complexity, quantity and sparsity [15][20].

MDE envisions components are present by DPIM (DICE Platform Independent Model) is first component which focus on data source and deploy the structure as nodes; the second component DTSM (DICE Technology Specific Model) afford combination of similar big data technology such as Hadoop, Spark, and NoSql with evaluation based on structure-view; the core construct component is DDSM (DICE Deployment Specific Model) mapped big data depend on cloud application; Model transformations component automatize earlier components via refinement, tradeoff, in place refactoring, and roll out [20].

3.0 DEVELOPMENT REFERENCES ARCHITECTURE

In this section modeling reference architecture. Subsequent, the significance use software concept in big data regard to cloud deployment.as aforementioned in section2.1 taxonomy of big data the mechanism of generate data source an important to determine the data type. While, Unstructured data request addition process to able acquisition information, the structure data forward to internal layers' data aggregation and analytics use cloud compute. The semi-structure data in mid-point which need to process as unstructured data.

Table 1 Constraints from data layer

source	Data Type	State	Request
Data Capture	Unstructured data	External	-Transformation - Preprocess - Storage into buffer to physical - Validate
Data Transformation	Semi-Structure	External/ Internal	-Transformation -Preprocess -Cloud (SaaS, Pass) -Validate
Data Consumption	Structure data	Internal	-Transformation -Cloud (Saas,Paas,Iaas) - Validate

Table.1, express the constraints from data layer are follows 1) transformation: drive to data from stages, components, and layer. 2) Validation: to ensure the data are storage. Big data dimensions 4Vs, the huge amount produced/generate data (volume).and type of data (variety) both are existing into data layer. Velocity dimension indicates the ability to response in action such transformation as storage which cloud application coverage uses the inner layers (Iaas, PaaS, and Iaas).

The two dimensions' veracity and value involving in data aggregation layer specifically acquisition, and analytics layer.

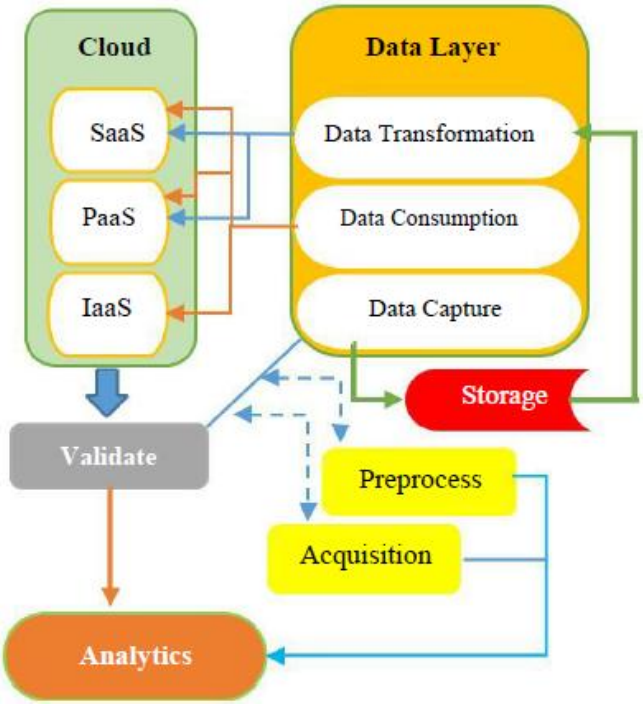

Figure 4 Development References Architecture

The figure above, illustrate the refinement process through components, which services oriented in cloud drive repository action from data layer to determine the source of raw data and type. An example the data capture is rolled out to datatransformation through storage, similar model transformations.

The act preprocess and acquisition are interlinking with data layer, validate and analytics layer, the aim of it to adjust the mapping data by two nodes content and physical. Meanwhile, allow the manipulate data from sources.

In cloud the SaaS play as API with data layer which conduct the authorities to data parallel with PaaS; IaaS the lower layer in cloud layer usage for internally which defined the storage capacity and indexing data regarding to life cycle of source data as well as the data storage in IaaS firm is structured and well organized.

Data capture request iteration in data layer to move the data digital data were captured externally into data transformation. In this point consider the buffer storage as limitation especially from big data concept,

Analytics layer is compatible with cloud and data layer regarding to validate aspect. For data aggregation layer consist of preprocess, storage and transformation.

4.0 CONCLUSIONS AND FUTURE WORK

Big data applications are significantly for academia, government, and IT companies; the securitize universal digital data encourage to research in big data wave, cloud computing develop fast to following the dimensions of big data.

We conclude based on previous work significant potential belong references architecture to combined various application in one platforms.in future work we plan to integrate layers and components through Services Oriented Architecture with Meta-Data Model.

REFERENCE

1. J. Gantz, "A forecast of worldwide information growth through 2010," *IDC (International Data Corp).*, 2007.
2. J. Gantz and D. Reinsel, "The Digital Universe in 2020: BigData,BiggerDigitalShadows,andBiggestGrowth in the Far East*," Int. Data Corp. (IDC)*., vol. 2007, no. December 2012, pp. 1–16, 2012.
3. PodkosovaandH.Kaufmann,"Mutualcollision avoidanceduringwalkinginrealandcollaborativevirtual environments," Proc. *ACM SIGGRAPH Symp. Interact. 3D Graph. Games* - I3D '18, pp. 1–9, 2018.
4. Economist, "Building with big data The data revolution is changing the landscape of business," *The economist*, 2016. [Online]. Available: https://www.economist.com/business/2011/05/26/building-with-big-data.
5. YouTube,"YouTubeinnumbers,"*YouTubePress*, 2018. [Online].Available: https://www.youtube.com/intl/en-GB/yt/about/press/.
6. R.Kune,P.K.Konugurthi,A.Agarwal,R.R. Chillarige, and R. Buyya, "The anatomy of big data computing," *Softw. - Pract. Exp.*, vol. 46, no. 1, pp. 79–105, 2016.
7. R. Grün, "A very personal, 35 years long journey in ESR dating," *Quat. Int.*, 2019.
8. X.Jin,B.W.Wah,X.Cheng,andY.Wang, "Significance and Challenges of Big Data Research," *Big Data Res.*, vol. 2, no. 2, pp. 59–64, 2015.
9. Microsoft,"ApplicationArchitectureCloudGuide,"2017. [Online]. Available: https://azure.microsoft.com/en-us/campaigns/cloud-application-architecture-guide/.
10. Google,"GoogleCloudPlatformOverview*,"https://cloud.google.com.* [Online]. Available:https://cloud.google.com/docs/overview/.
11. Oracle, "Oracle Cloud Infrastructure Documentation,"2018. [Online]. Available: https://docs.cloud.oracle.com/iaas/Content/Block/Concepts/overview.htm.
12. R. G. Adam Ronthal, RoxaneEdjlali, "Magic Quadrant for Data Management Solutions for Analytics," 2019.
13. P. Pääkkönen and D. Pakkala, "Reference Architecture and Classification of Technologies, Products and Services for Big Data Systems," *Big Data Res.*, vol. 2, no. 4, pp. 166–186, 2015.
14. F. Liu *et al*., "Rudall et al. 1995."
15. M. N. Zafar, F. Azam, S. Rehman, and M. W. Anwar, "A Systematic Review of Big Data Analytics Using Model Driven Engineering," pp. 1–5, 2017.
16. S. FossoWamba, S. Akter, A. Edwards, G. Chopin, and D. Gnanzou, "How 'big data' can make big impact: Findings from a systematic review and a longitudinal case study," *Int. J. Prod. Econ.*, vol. 165, pp. 234–246, 2015.
17. Y. Wang, L. A. Kung, and T. A. Byrd, "Big data analytics: Understanding its capabilities and potential benefits for healthcare organizations," *Technol. Forecast. Soc. Change*, vol. 126, pp. 3–13, 2018.
18. K.SchmidtandI.Wagner,"Orderingsystems: Coordinative practices and artifacts in architectural design and planning," *Comput. Support. Coop. Work*, vol. 13, no. 5–6, pp. 349–408, 2004.
19. D. D. Cloud, "DICE Methodology," 2017.
M. Guerriero, S. Tajfar, D. A. Tamburri, and E. Di Nitto, "Towards a model-driven design tool for big data architectures," pp. 37–43, 2016.