Big Data Challenges for E-governess System in Distributing Systems

M. Suresh, R. Parthasarathy, M. Prabakaran, S.Raja

***Abstract*—This paper discusses the challenges that are imposed by E-governess on the modern and future infrastructure. This paper refers to map reducing algorithm to define the requirements on data management, access control and security. This model that includes all the major stages and reflect specifying data management in modern E-government. This paper proposes the map reducing architecture model that provides the basic for building interoperable data. The paper explain how the implemented using Distributed structures and provisioning model.**

Index Terms —Big Data, Map reducing, Distributed structures.

I. INTRODUCTION

Modern infrastructure allows targeting new large scale problem-governesses which solution was not possible before ,e.g. Banking, Media, Airlines, Telecom, Entertainment News, Sports, Astrology, Movie Tickets, Public Works Monitoring, Electricity Board, Health etc. e-governess typically produces a huge amount of data that need to be supported by a new type of e-Infrastructure capable to store, distribute, process, preserve, and curate these data:

We refer to these new infrastructures as E-governess Data Infrastructure. In e-governess, the data are complex multifaceted objects with

the complex internal relations, they are becoming an infrastructure of their own and need to be supported by corresponding physical or logical infrastructures to store, access and manage these data.

The emerging E-GOVERNESS should allow different groups of researchers to work on the same data sets, constructing their own distributed approach and collaborative environments, safely store and retrieved intermediate results, and later share the discovered results. New data provide, Third party security and access control mechanisms and tools will allow researchers to link their e-governess results with the initial data and intermediate data to allow future re-use/re-purpose of big data, e.g. with the improved research technique and tools.

The paper is organized as follows. Section 2 gives an overview of the main research communities and summarizes requirements to future E-GOVERNESS. Section 3 discusses challenges to data engagement in Big Data E-governess, including Map reduce discussion. Section 4 introduces the proposed E-governess architecture model that is intended to answer the future big data challenges and requirements. Section 5 discusses e-governess implementation using Distributed technologies.

Manuscript Received on November, 2013.

M,Suresh,, Computer Science Engineering- Muthayammal Engineering College-India.

R.Parthasarathy Computer Science Engineering- Muthayammal Engineering College-India.

M,suresh,, Computer Science Engineering- Muthayammal Engineering College-India.

S.Raja,, Computer Science Engineering- Muthayammal Engineering College-India.

II. GENERAL REQUIREMENTS TO BIG DATA E-GOVERNESS

Big Data e- governess is becoming a new technology driver and requires re-thinking a number of infrastructure components, solutions and processes to address the following general challenges:

- Exponential growth of data volume produced by different research instruments and/or collected from sensors
- Need to consolidate e-Infrastructure as persistent research platform to ensure research continuity and cross-disciplinary collaboration, deliver/offer persistent services, with adequate governance model. The recent advancements in the general ICT and big data technologies facilitate the paradigm change in modern e-governess.

E-governess that is characterized by the following features:

- Automation of all e- governess processes including data collection, storing, classification, indexing and other components of the general data duration and provenance Transformation all processes, events and products into digital form by means of multi-dimensional multifaceted measurements, monitoring and control; digitising existing artefacts and other content.
- Possibility to re-use the initial and published E-governess with possible data re-purposing for secondary research
- Global data availability and access over network for cooperative group of public user including wide public access to e-governess
- Advanced security and access control technologies that ensure secure operation of the complex research infrastructures and e-governess instruments and allow creating trusted secure environment for cooperating groups and individual researchers
- Multidimensional data can involved and distributing the data management efficiently
- Monitoring the heterogeneous data might be evaluated constructing in the structure manner
- The data management efficiently could be stored and retrieved by using the big data much effectively in corresponding technology

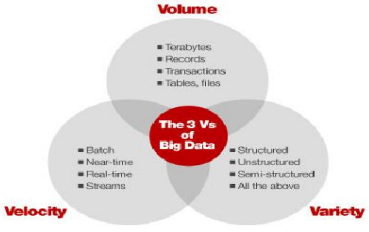

Figure: Big data structures

A) E-governess Requirements in Research communities

Informatics has emerged as the thrust area for the Government as it can enable the administration to re-engineer and improve its processes, connect citizens and build interactions with and within the society by ringing radical changes in its functioning leading to **S**imple, **M**oral, **A**ccountable, **R**esponsive and **T**ransparent **(SMART)** governance. This new way of governance adopted by the public administration for the delivery of services on the Internet and Intranet, constitute the concept of Electronic Governance (E-Governance).

E-Governance to change potentials for sufficient development:

- Automation replacing existing manual processes, which involve accepting, storing, processing, outputting or transmitting data/information was more efficient
- Corresponding information supporting current processes of decision-making and implementation.
- Transmission of information by extensive use of Electronic Forms and Interfaces through use of web and Internet technology (Paperless Government-On-Line).

With the availability of information technologies including Database Management, Data Warehousing and Data Mining, e-Governance is aimed at converting the transactional data into business relevant information and managing this repository. This asset, in turn, is made accessible to the decision makers by computer based Decision Support System (DSS) using alphanumeric information in various application areas.

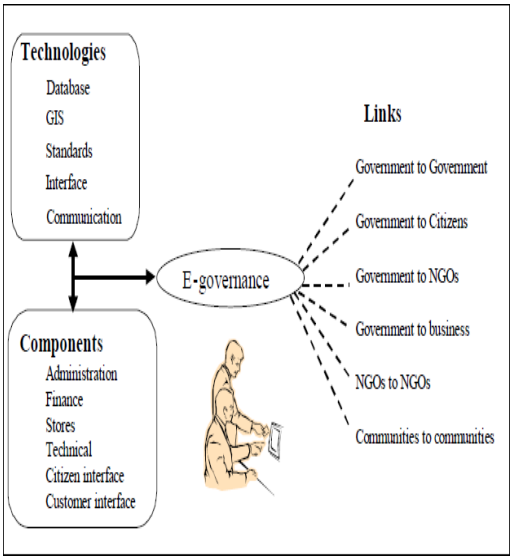

Figure: E-governess data Management structures

The infrastructure requirements to E-GOVERNESS for emerging Big Data E-governess:

- High Volume of data supported very long time
- Large Volumes of generated data at high speed
- Multi dimensional data distribution and replication
- Support of virtual e-governess communities
- Security environment for data storage and retrieval processing
- Data integrity, confidentiality, accountability
- Binding the privacy by policy

Distributed computing

Distributed computing is a technique to optimize use of resources over the networking. It is a way to increase the capacity or add capabilities dynamically without investing in new infrastructure, training new personnel, or licensing new software. It extends Information Technology's (IT) existing capabilities. Distributed computing entrusts remote services with a user's data, software and computation. Moving data into the Distributed offers great convenience to users since they don't have to care about the complexities of direct hardware management. For a quality service data security is a necessary thing, security must be imposed on data by using encryption strategies to achieve secured data storage and access. The transparent nature of Distributed it is necessary to anxious about the security issues. But distributed infrastructure even more reliable and powerful then personal computing, but wide range of internal, external threats for data stored on the distributed system. Since the data are not stored in client area, implementing security measures cannot be applied directly on e - governance.

We can use the Distributed computing on every field of e – Governance.

- Government to Citizen (G2C)
- Government to Government (G2G)
- Government to Business (G2B)
- Government to Enterprise (G2E)
- Government to NGO (G2N)

III. DATA MANAGEMENT IN BIG DATA AN E-GOVERNESS

Emergence of computer aided research methods is transforming the way how research are done and e-governess data are used. The following types of e-governess data are defined [4]:

- Raw data collected from observation and from Experiment (according to an initial research model)
- Structured data and datasets that went through data Filtering and processing (supporting some particular formal model)
- Published data that supports one or another e-governess hypothesis, research result or statement Data linked to publications to support the wide research consolidation, integration, and openness.

Volume refers to larger amounts of data being Generated from a range of sources For example, big data can include data gathered from the Internet of Things (Iota). As originally conceived, 3 Iota referred to the data gathered from a range of devices and sensors networked together, over the Internet. RFID tags appear on inventory items capturing transaction data as goods are shipped through the supply chain. Big data can also refer to the exploding information available on social Media such as Face book and Twitter.

Variety refers to using multiple kinds of data to Analyze a situation or event. On the Iota, millions of devices generating a constant flow of data results in not only a large volume of data but different types of data characteristic of different situations. For example, in addition to WSN, heart monitors in patients and Global position System all generate different types of structured data However,

devices and sensors aren't the only sources of data. Additionally, people on the Internet generate a highly diverse set of structured and unstructured data. Web browsing data, captured as a sequence of clicks, is structured data. However, there's also substantial unstructured data. For example, according to kingdom, 4 in 2012 there were 600 million websites and more than 125 million blogs, with many including non structured multidimensional data base..
As a result, there's an assemblage of data emerging through the "Internet of People and Things"5 and the "Internet of Everything."
Velocity of data also is on demand rapidly over time for semi structure data band there's a need for more frequent decision making about that data. As the world becomes more global and developed, and as the Iota builds, there's an increasing frequency of data capture and decision making about those "things" as they move through the world. Further, the velocity of social media use is increasing. For example, there are more than 250 million face book per day.4 Face book lead to decisions about other Face book, escalating the velocity. Further, unlike classic data warehouses that generally "store" data, big data is more dynamic. As decisions are made using big data, those decisions ultimately can influence the next data that's gathered and analyzed, adding another dimension to velocity.

IV. MAPREDUCE AND HADOOP

Map Reduce has been used by Google, facebook, amazon, yahoo etc to generate scalable applications. Inspired by the "map" and "reduce" functions in Lisp, Map Reduce breaks an application into several small portions of the problem, each of which can be executed across any node in a computer cluster. The "map" stage gives sub problems to nodes of computers, and the "reduce" combines the results from all of those different sub problems. Map Reduce provides an interface that allows distributed Computing and parallelization on clusters of computers Map Reduce is used at Google, facebook,amazon,yahoo etc for a large number of activities, including data mining and machine learning. Hadoop (http://hadoop.apache.org), named after a boy's toy elephant, is an open source version of Map Reduce. Apparently,
Facebookhttp://developer.facebook.com/hadoop) is the largest user (developer and tester) of Hadoop,with more than 550 million users per month and billions of transactions per day using multiple pet bytes of data.7 As an example of the use of the Map Reduce approach, consider a Face book front page that might be broken into multiple categories—such as advertisements (optimized for the user), must-see videos (subject to content optimization), news (subject to content management), and so on—where each category could be handled by different clusters of computers. Further, within each of those areas, problems might be further decomposed; facilitating even faster response.Map Reduce allows the development of approaches that can handle larger volumes of data using larger numbers of processors. As a result, some of the issues caused by increasing volumes and velocities of data can be addressed using parallel-based approaches.

- Reduced the complexity
- More Effectiveness
- Robustness
- Scalable and Elastics
- Wide range of application

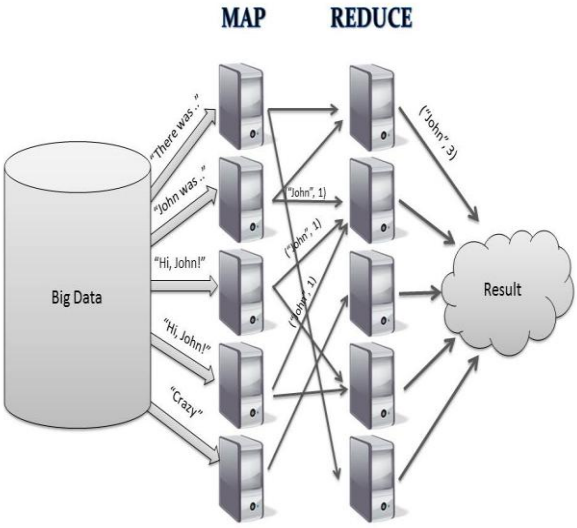

Figure: map reducing algorithm

V. DISTRIBUTED VIRTUAL TECHNOLOGY REQUIREMENTS

Technologies for a distributed solution for a comprehensive e-governance solution, that meets the Objectives defined in the earlier sections, will have to Address many diverse requirements that may be present due to various reasons. These reasons may be economic, political, technical and cultural amongst others. The requirements are classified into two categories, (a) distributed technology requirements that discusses the core technology requirements and (b) application requirements, which discusses abstraction of common code required for multiple applications/ departments
VM technologies' increasing ubiquity has enabled users to create customized environments atop physical infrastructure and has facilitated the emergence of business models such as Distributed computing. VMs' use has several benefits:

- Server consolidation, which lets system administrators place the workloads of several underutilized servers in fewer machines; the ability to create VMs to run legacy code without interfering with other applications' APIs;
- improved security through the creation of sandboxes for running applications with questionable reliability; and
- Performance isolation, letting providers offer some guarantees and better quality of service to customers' applications.

Existing VM-based resource management systems can manage a cluster of computers within a site, allowing users to create virtual workspaces5 or clusters.6 Such systems can bind resources to virtual clusters or workspaces according to a user's demand. They commonly provide an interface through which users can allocate VMs and configure them with a chosen operating system and software. These resource managers, or *virtual infrastructure engines* (VIEs), let users create customized virtual clusters by using shares of the physical machines available at the site.

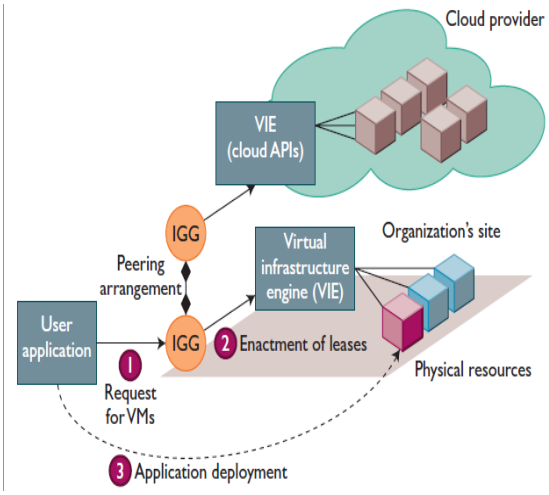

Figure: Distributed Virtual Infrastructures

Distributed Computing Services

- ***Distributed Services Delivery Platform:***
  This is essentially a workflow engine that executes the application which - as we described in the previous section, is ideally composed as business workflow that orchestrates a number of distributable workflow elements. This defines the services dial tone in our reference architecture model.
- ***Distributed Services Creation Platform:***
  This layer provides the tools that developers will use to create applications defined as collection of services which can be composed, decomposed and distributed on the fly to virtual servers that are automatically created and managed by the distributed services assurance platform.

Big Data Security System as Distributed Computing Advantages

There are many points for database security system as follows

A. Reduced Costs

Today we know that the Business and IT leaders understand the need for accurate and timely information when making decisions that impact their business. And calculate the accurate figure to our business aspects. To ensure that the right information is available when it is needed, IT projects often spend a large portion of their time, resources and money to create what amounts to individual information silos, with distinct requirements, configurations, and support models. Historically, production environments have been configured for peak load, peak performance and uninterrupted business continuity.

B. Distributed Computing Framework Increase the Service Levels

In these fields the service orientation aspects of DBaaS architectures benefit both IT providers and consumers. Providers benefit from being able to develop and offer pre-defined services for their consumers to use – minimizing vendor, software version and configuration diversity. This reduced diversity supports business goals of agility, efficiency and improved quality of service through the development of standardized processes, common support mechanisms and focused skills development.

C. Access the Enhanced Information to server to client

Another common practice within organizations stems from the misperception that information requirements are so unique that each line of business or region must maintain separate and dedicated database environments.

D.Distributed Big Data Storage Model

In the Distributed Computing System end user always store there data in Distributed not in their local system. Then it's must that distributed computing Give a effectiveness and correct Secure Data in distributed Distributed Storage. It's possible that an unauthorized person modify the data or access data the distributed case when such inconsistencies are successfully detected, to find which server the data error lies in is also of great significance, since it can be the first step to fast recover the storage errors.

The homomorphic token is introduced. The token computation function we are considering belongs to a family of universal hash function. It is also shown how to derive a challenge response protocol for verifying the storage correctness as well as identifying misbehaving servers. Finally, the procedure for file retrieval and error recovery based on erasure-correcting code is outlined.

VI. FUTURE RESEARCH AND DEVELOPMENT

The future research and development will include further SDLM definition, e-governess and Map Reduce components definition and development with focus on infrastructure components of e-governess. Special attention will be given to defining the whole cycle of the provisioning e-governess services on-demand specifically tailored to support instant e-governess workflows using Distributed platforms. This research will be also supported by development of the corresponding Big Data e-governess processes and Map reduces operation.

REFERENCES

1. Bansal, V. and J. Bhattacharya. E-governance solution for government of Maharashtra. Technology whitepaper, India Research Lab, IBM, 2000.
2. Batra, V.; J. Bhattacharya; H. Chauhan; A. Gupta; M.Mohania; U. Sharma. 2002. "Policy Driven Data.
3. Administration". In POLICY 2002, IEEE 3rd International Workshop on Policies for Distributed Systems and Networks.
4. Gillick et al., 2006, Gillick D., Faria A., DeNero J., and MapReduce: Distributed Computing for Machine Learning, Berkley, and December 18, 2006.
5. K. Shvachko, Hairong Kuang, S. Radia and R Chansler – The Hadoop Distributed File System. Mass Storage Systems and Technologies (MSST), 2010 IEEE 26th Symposium 3-7 May 2010.
6. F. Chang, J. Dean, S. Ghemawat, W. Hsieh, D. Wallach,M. Burrows, T. Chandra, A. Fikes, and R. Gruber, "Bigtable: A distributed structured data storage system," in 7th OSDI,2006, pp. 305–314.
7. A. Stupar, S. Michel, and R. Schenkel, "Rankreduce–processing k-nearest neighbor queries on top of mapreduce," in Proceedings of the 8th Workshop on Large-Scale Distributed Systems for Information Retrieval, 2010, pp. 13–18.
8. J. Ekanayake, H. Li, B. Zhang, T. Gunarathne, S. Bae, J. Qiu, and G. Fox, "Twister: a runtime for iterative mapreduce," in Proceedings of the 19th ACM International Symposium on High Performance Distributed Computing. ACM, 2010, pp. 810–818.
9. J.S. Chase et al., "Dynamic Virtual Clusters in a Grid Site Manager," Proc. 12th IEEE Int'l Symp. High Performance Distributed Computing (HPDC 03), IEEE CS Press, 2003, p. 90.