Big Data Migration in the Cloud Environment - An Overview

V.Rathika
Assistant Professor in Computer Applications
Idhaya College for Women, Kumbakonam
Tamil Nadu,India

L. Arockiam, Ph.D.
Associate Professor in Computer Science
St. Joseph's College, Trichy
Tamil Nadu,India

ABSTRACT
Big data migration refers moving of large quantity of data from source to destination. Based on population or digitized usages this era requires new technologies to handle large volume of data. So we have to move from legacy system to new system. As well as this migration will reduce cycle times, lower costs of maintenance, increase access to critical data, improve software architecture, improve scalability and reliability. All systems in the organizations may be replaced or enhanced the functionality with other systems. Data migration is very important for a variety of reasons such as server to storage equipment replacements or upgrades, website consolidation, server maintenance and data center relocation. ETL tools are supported the data migration which is the business process but also a set of technical applications and people's knowledge. But this move is not easy to achieve. It has more challenges and issues. This paper is going to provide the important concepts of big data migration.

General Terms
Big Data Migration, Cloud Computing

Keywords
Cloud Computing, Big Data, Migration, Process Model, Types, Issues and Challenges, Methodology, and Technologies.

1. INTRODUCTION
The process of transferring large volume of data between computer systems, storage types and formats is called big data migration. It is recommended when Organizations upgrade the systems for their needs or replace their computer systems. In migration, data has to be transferred from source to target [1]. Effective data migration procedure has design, data extraction, data loading, data verification and data cleansing as phases to accomplish successful migration. Design can create phases such as data extraction and data loading. Data extraction defines where data is read from the old system. And data loading defines where data is written to the new system. Data verification occurs after loading which determines whether data was accurately translated. Data cleansing is performed automatically to improve data quality, eliminate redundant or obsolete information [18].

2. CLOUD COMPUTING
Fifth generation of computing is called Cloud Computing [22]. It is the fastest growing part of IT. Cloud services are simpler to acquire and scale up or down. Reduced cost, flexibility, improved automation, focus on core competency and sustainability are some of the benefits of cloud computing.

Cloud deployment model has Public, Private and Hybrid clouds. Public clouds are run by third party companies such as Google, Amazon and Microsoft. It offers services to multiple customers over a common infrastructure. Private clouds are dedicated a single organization which are designed by third party. It offers services through private networks. Hybrid clouds are the combination of private and public clouds. Noncore applications accessed by organizations in a public cloud and core applications are maintained in the private cloud.

Cloud Service delivery model has three layers such as Software as a Service (SaaS), Platform as a Service (PaaS) and Infrastructure as a Service (IaaS). SaaS offers application services over the network on a subscription and on – demand basis. PaaS offers runtime environment, Software development framework over the network on a Pay as you go basis. IaaS offers computing services, entire infrastructure to run customer applications and storage facility over the network on pay as you go basis.

Security, difficulty to migrate, costing model, charging model, cloud interoperability issue and authentication are the challenges in the cloud computing [23].

3. BIG DATA ANALYTICS
Big data [24] is the collection of both structured, semi structured, and unstructured data which are from different sources like social data, machine generated data and traditional enterprises. Large volume of data is generated from mobile devices and big companies such as Google, Apple, Face book, Yahoo and Twitters.

Big data has five characteristics are Volume, Velocity, Variety, Veracity and Value. Volume refers to the vast amount of data. Velocity refers to speed rate to collect or acquire or generate or process of data. Variety refers to different types of data and data sources. Veracity refers to the quality and reliability of the data. Value determines valid data which increases business value to gain advantages.

Apache Hadoop, Map Reduce, Oracle big data appliance, Parallel DBMS technologies is some of the tools available to handle big data. Capturing of big data, storage, data analysis, retrieval and lack of structure are the main challenges [25] associated with big data. Big data is used in medicine, astrometry, biology, simulation RFID, physics, and healthcare monitoring, manufacturing and transportation management.

4. RELATED WORK
In [10], Atal Srivastav explained about Pre-Migration and Post-Migration Architecture in his "Data migration - Case Study". He collected client's problem like High operating costs, large number of inconsistent data, and slow speed and Compatibility issues between databases. He utilized technologies like Oracle Database (10g, 11g), SQL Server 2008, .net Framework 3.5 and WSDL 2.0/XML1.0/Web Services (SOA) to rectify customer's problems.

In [11], Brodie and Stonebraker proposed an 11 step Chicken Little methodology for migration. Here, Legacy and target systems are operated in parallel throughout the migration process. It provides more mature approach. But it doesn't provide any guidelines for testing the process. Bing et al. proposed Butterfly methodology for migration. It moves a mission – critical legacy system to a target system in a simple, fast and safe way. It eliminates complexity of maintaining the consistency between source and target.

In [19], E. Anderson et al. considered the problem of finding an efficient migration plan. They focused on offline migration only which can be performed as a background process or at a time when loads from user requests are low. Their goal was to find a migration plan that uses the existing network connections between storage devices to move the data from the initial configuration to the final configuration in the minimum amount of time. They proposed two new algorithms such as Max Degree Matching and Greedy Matching algorithm.

In [21], Nasuni compared transfer rates among three cloud service providers such as Amazon S3, Microsoft Windows Azure and Rackspace. It considered 12 TB volume of data to move. Based on this, it reported the result as 40 hrs need from S3 to Azure, 4 hrs need from Azure to S3, 5 hrs need from Rackspace to S3, 7 days need from S3 to Rackspace and 4 hrs need from one S3 to another S3.

5. MIGRATION PROCESS

Data migration process (see Figure 1) has 4 steps. They are Evaluate, Review, Restore and Migrate [2][13]. Here, Inaccessible data refers to all brands and media. Accessible data refers Streamlined Data Methodologies, Proactive Management, Reduce Costs and Retention Policies.

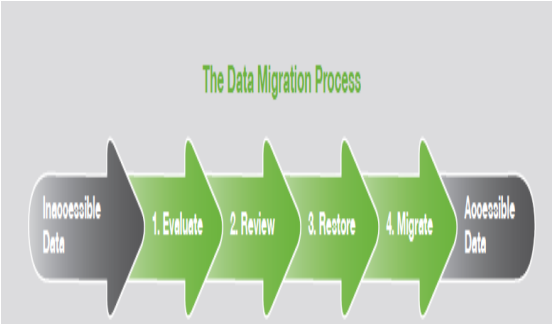

Figure 1 – Data Migration Process

Step1 - Evaluate – It evaluates Media type, State of Media and estimated effort.

Step2 - Review – It reviews Customer Requirements and criteria.

Step3 - Restore – It re-establishes Recovery, Extraction and Discovery.

Step4 - Migrate – It refers to the transformation of data and purges unnecessary files.

6. DATA PROCESS MODEL

Data process model (see Figure 2) has four phases. It explains the logical and temporal dependencies between the tasks [3]. As well as it clarifies roles and responsibilities in a migration project. It has stability and reliability.

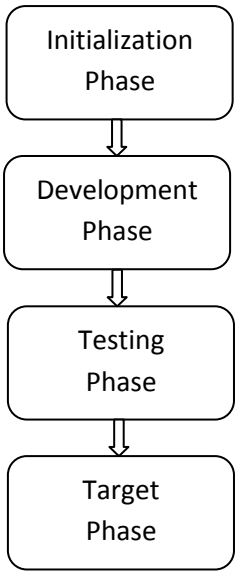

Figure 2 – Data Process Model

6.1 Initialization Phase

This phase refers the tasks Call for Tender, Strategy & Pre analysis and Platform Setup. It collects Customers queries and conducts pre analysis to find strategy. Finally it creates migration platform and infrastructure.

6.2 Development Phase

This phase consists of source and target and structure analysis, source data cleansing and data transformation. It maintains data quality while transmitting data from source to target.

6.3 Testing Phase

This phase has Migration run test, Appearance test, and Completeness & Type correspondent test, Process ability test, Integration test and final rehearsal. It tests whether all data are migrated completely and correctly.

6.4 Target Phase

After the completion of successful migration, target system is up and running while source is shut down.

7. TYPES OF MIGRATION

7.1 Database Migration

Database Migration refers data are moving from one Database to another. It means that data are extracted from one database and loaded into another without any data loss. It should be between two database engines such as from SQL server to Oracle or from old version to new version[6].

7.2 Application Migration and Business Process Migration

Both Migrations are more or less similar. In application migration, Data are moved from one application to another when a company decides to change software.

In Business process migrations, entire processes are moved from one process to another process. It is highly complicated one [9].

7.3 Storage Migration

Storage Migration refers to data will be moved from old and outdated storage place to a new one.

8. MIGRATION METHODOLOGY

Dell introduced this methodology (see Figure 3) for data migration. It has five steps to implement successful data migrations [16].

Figure 3 – Data Migration Methodology

8.1 Planning

This planning creates Migration Plan. It defines Migration goal, requirements and HW/SW tools. It concentrates on methodology also.

8.2 Pre – Production Testing

This stage arranges testing migration environment. It validates data, HW/SW and migration tools. It performs migration testing based on plan and updates final migration plan.

8.3 Migration

This creates final Migration environment and installs migration software. After that it performs migration based plan.

8.4 Validation

This validation collects migration statistics. It verifies the completion of the migration. It validates application functionality and creates migration report.

8.5 Cutover

This moves application to the target storage environment based on cutover plan. It runs the new environment during the defined time frame. It creates final report and retires the previous environment.

9. MIGRATION ALGORITHMS

There are number of algorithms [19] available for data migration as follows

- The 2 – factoring algorithm
- 2- factoring a multigraph
- The bypass algorithm for 4 regular multigraphs
- The reduction to 4 –regular graphs
- Max-degree matching
- Greedy matching

This paper concentrates on Max-Degree Matching and Greedy Matching algorithms.

9.1 Max – Degree Matching Algorithm

9.1.1 Definition

This algorithm (see 6.1.2 Algorithm) uses 2n/3 bypass nodes and has Δ step migration plan without space constraints. It works by sending, in each stage, one object from each node in the demand graph that has maximum degree.

9.1.2 Algorithm

1. Set up a bipartite matching problem as follows: the left hand side of the graph is all maximum degree vertices not adjacent to degree one vertices in G, the right hand side is all their neighbors in G, and the edges are all edges between maximum degree vertices and their neighbors in G.
2. Find the maximum bipartite matching. This solution induces cycles and paths in the demand graph. All cycles contain only maximum degree vertices; all paths have one endpoint that is not a maximum degree node.
3. Mark every other edge in the cycles and paths. For odd length cycles, one node will be left with no marked edges. Make sure that this is a node with an outgoing edge (and thus can be bypassed if needed). Each node has at most one edge marked. Mark every edge between a maximum degree node and a degree one node.
4. Let V' be the set of vertices incident to a marked edge. Compute a maximum matching in G that matches all vertices in V'. Define S to be all edges in the matching.
5. For each node u of maximum degree with no incident edge in S, let (u,v) be some out-edge from u, Add (u,b) to S, where b is an unsued bypass node, and add(b,v) to the demand graph G.
6. Schedule all edges in S to be sent in the next stage and remove these edges from the demand graph.
7. If there are still edges in the demand graph, go back to step 1.

9.1.3 Theorem

Max – Degree – Matching algorithm computes a correct Δ stage migration plan using at most 2n/3 bypass nodes.

9.1.4 Proof

First it shows that the algorithm uses no more than Δ stages. Hall's theorem shows that the matching problem which is described in Step 1 has a solution in which all the maximum degree vertices are matched. Maximum degree node is decreased by one on each stage. So, after Δ stages there is no edges left and are done.

Second it shows that the algorithm uses no more than 2n/3 bypass nodes. Let k be the number of paths and cycles in G. Each path has at least two vertices, so $k \leq n/2$. n_b denotes number of bypass nodes of previous stage. n_m denotes number of maximum degree vertices without bypass nodes. $n_m /3$ denotes number of bypass nodes of current stage. So, maximum number of bypass nodes will be received from the following equation: $n_m /3 + n_b \leq (n-n_b) /3 + n_b = n/3 + 2n_b /3$. It proved that only 2n/3 bypass nodes are required for entire process.

9.2 Greedy Matching Algorithm

9.2.1 Definition

This algorithm (see 6.2.2 Algorithm) is a straight forward direct migration algorithm with space constraints. It sends all objects.

9.2.2 Algorithm

1. Let G' be the graph induced by the sendable edges in the demand graph. An edge is sendable if there is free space at its destination.
2. Compute a maximum general matching on G'.
3. Schedule all edges in matching to be sent in this stage.
4. Remove these edges from the demand graph.
5. Repeat until the demand graph has no more edges.

9.2.3 Lemma

Given initial free space of $fi \geq 1 + \max(0, d_{in}(vi) - d_{out}(vi))$, at any stage of a direct migration (after sending any number of objects) at least one unsent object is sendable.

Where v denotes node, $d_{in}(v)$ denotes in-degree of the node and $d_{out}(v)$ denotes out-degree of the node.

9.2.4 Proof

It shows that an each stage one unsent node is sendable to destination. $v_* \in V'$ denotes exist node which has , $d_{in}(v_*) - d_{out}(v_*) \geq 0$. So, it has free space and an incoming edge which is sendable.

10. ISSUES AND CHALLEGES

10.1 Business Factors and Technical Factors

- Existing Investments in IT.
- Costs and Data Security.
- Regulations and Provisioning.
- Existing infrastructure.
- Security Architecture.
- Complexity.
- Network and support.
- IT Skills.
- Service Level Agreements are Technical Factors [4].

10.2 Software and Hardware Issues

- License renewals and extensions are software issues.
- Redundancy and disposal are hardware issues [15].

10.3 Run Time Issues

- Number of duplicate.
- Incorrect and inconsistent records.
- Minimal Data model changes permitted multiple source databases and data models.
- huge difference between data models at source and data models at target [3].

10.4 Some Common Challenges

- Lack of available expertise.
- Poorly executed scoping and budgetary analysis.
- Lack of data quality management strategy and appropriate tools.
- Lack of documentation and detailed knowledge of legacy and target environments.
- Data Migration methodology is insufficient or ignored completely.
- Lack of collaboration between cross – party project teams.
- Poor choice of data migration technology, Project delivery approach is inflexible.
- Go – Live Strategy is inappropriate for the needs of the business and Target application is constantly changing [8].

10.5 Five pitfalls

- Failing to engage the lines of business and business users at the outset.
- Absence of data governance policies and organizational structure.
- Poor data quality in legacy system.
- Neglecting to validate and redefine business rules and Failure to validate and test the data migration process [8][12].

10.6 Five Big Risks

- Employees entrusted with a data migration project lack industry best practices experience.
- Team relies too much on the tools of the job.
- Cross object dependencies.
- Attempting to go live in one big upload at the end and Budget overruns due to inadequate scoping or preparation at the start [14].

11. SOME SOLUTIONS FOR ENSURING SUCCESS

There are some solutions to the above mentioned common challenges.

- Start Early.
- Obtain adequate resources.
- Train Early.
- Integrate the Schedule.
- Ensure Cooperation.
- Use Professional.
- Impose Discipline.
- Manage Changes and Manage risks [8].

12. MIGRATION TECHNOLOGIES

There are four technologies [5] [7] for migration.

12.1 ETL (Extract, Transform, Load)

It handles terabyte scale datasets and multi pass transformations. It also has deep data profiling, interoperability with data quality and many to many data integration capabilities.

12.2 Hand Coding

Developers can create coding for an economic superiority tools to do math and recognize.

12.3 Replication

It has high end and low end replication tools. Low end replication tools are allowing data to move one way only. High end replication tools are having bidirectional, transformational, heterogeneous data synchronization capabilities.

12.4 EIA (Enterprise Application Integration)

It handles small amount of data efficiently and quickly. But it couldn't handle extreme volume, data quality and profiling.

13. MIGRATION ARCHITECTURE

This architecture [20] proposed by Microsoft for migration (see Figure 4). Source entity explains the data and its format in the source. All these data are moved to Staging. Staging has tables to map the source data. Data are cleansed here and prepared for migration.

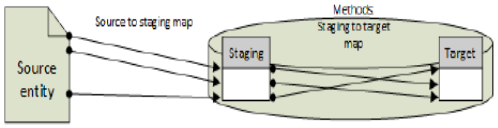

Figure 4 – Data Migration Architecture

After this, Data migrated to target map. In this stage data are validated by Microsoft's method. At last data moved to target.

14. MIGRATION STRATEGIES

There are two types of strategies that are Big Bang Migrations and Trickle Migrations [17]. These are depending on the project requirements and available processing windows.

14.1 Big Bang Migration

It completes the entire migration in a small, defined processing window. In this, Data are extracted from the source systems, processed, and loaded to the target, followed by the switching of processing over to the new environment.

It offers off line migrations. It means, source is shutdown while migration in process. It completes migration in the shortest time but it has several risks.

14.2 Trickle Migration

It completes the entire migration in a short time window. It involves running the old and new systems in parallel and migrate the data in phases. It offers real time processes to move data.

15. CONCLUSION

Migration is important factor when new systems are launched. Varieties of strategy, technology and methodologies are available for Migration. Even though, ETL is popular technology to handle migration. It matches unstructured data with structured data. It integrates with data quality tools and many incorporate tools for data cleansing and mapping.

16. REFERENCES

[1] Rashmi, Dr.Shabana Mehfuz, Dr.G.Sahoo. 2012. A Five Phased Approach for the Cloud Migration. International Journal of Emerging Technology and Advanced Engineering, Volume 2, Issue 4.

[2] Andreas R uping. 2010. Transform!-Patterns for data migration. Proceedings of the 15th European Conference on Pattern Languages of Programming.

[3] Philip Russom. 2006. Best Practices in Data Migration. TDWI (Informatica).

[4] Jeff Bertolucci. 2012. 10 Big Data Migration Mistakes. Information Week.

[5] Cisco Data Center. 2011. Cisco Data Center Migration Service. Cisco Public Information.

[6] Klaus Haller. 2009. Towards the industrialization of data migration: concepts and patterns for standard software implementation projects. Proceedings of CAISE.

[7] Ch.Sai Krishna Manohar. 2013. A Greener Approach to Cloud Computing using Virtual Migration. International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 8.

[8] Dylan Jones. 2009. Some Common Data Migration risks (and how to avoid them). Data Migration.

[9] Klaus Haller.2008. Data Migration project management and standard software experiences in analog implementation projects. Proceedings of DW2008 Conference.

[10] Atul Srivastav. 2010. Data Migration - Case Study. www.enterprise data migration.com.

[11] Bing Wu, Deirdre Lawless, Jesus Bisbal, Jane Grimson. 1997. Lagacy System Migration: A Legacy Data Migration Engine. 17th International Database Conference.

[12] Rajkumar Buyya, Anton Beloglazov, Jemal Abawajy. 2009. Energy efficient management of data centre resources for cloud computing: A vision, architectural elements and open challenges. 9th International Symposium on Cluster Computing.

[13] Martin Wagner, Tim Wellhausen. 2011. Patterns for Data Migration Projects. www.TNGTECH.com.

[14] A. Khajeh-Hosseini, I.Sommerville,and D.Greenwood. 2010. Cloud Migration: A Case Study of Migrating an Enterprise IT System to IaaS. 1st ACM Symposium on Cloud Computing.

[15] P.Mohagheghi, T.Saether. 2010. Software Engineering Challenges for Migration to the service cloud Paradigm Ongoing Work in the REMICS Project. IEEE World Congress on Services.

[16] Dell's White Paper. 2012. Methodologies for Data Migration to Dell Fluid Data Architecture.

[17] An Oracle's White Paper. 2011. Successful Data Migration. www.oracle.com.

[18] Kam Woods, Geoffrey Brown. 2008. Migration performance for legacy Data Access. The International Journal of Digital Curation, Issue 2, Volume 3.

[19] E.Anderson, J.Hall, J.Hartline, M.Hobbes, A.Karlin, J.Saia, R.Swaminathan and J.Wilkes. 2008. Algorithms for Data Migration. Springer.

[20] Microsoft's White Paper. 2014. Data Import / Export framework user guide.

[21] Nasuni's White Paper. 2012. Bulk Data Migration in the Cloud. www.nasuni.com.

[22] Bhustan Lalsahu, Rajesh Tiwari. 2012. A Comprehensive Study on Cloud Computing. International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, Issue 9.

[23] P.Radha Krishna Reddy, S.Pavan Kumar Reddy, G.Sireesha and U.Seshadri. 2012. The Security Issues of Cloud Computing overall Normal & IT Sector. International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, Issue 3.

[24] Anuradha Bhatia and Gaurav Vaswani. 2013. Big Data – An Overview. International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, Issue 8.

[25] C.Bizer, P.Bonez, M.L.Bordie and O.Erling. 2011. The meaningful use of Big Data : Four Perspective – Four Challenges. SIGMOD, Vol. 40, No.4.