IOSR Journal of Business and Management (IOSR-JBM)
e-ISSN: 2278-487X, p-ISSN: 2319-7668. Volume 20, Issue 6. Ver. VI. (June. 2018), PP 01-12
www.iosrjournals.org

Taming Tsunami of Data by Principles of Inventory Management

Sarvesh Kumar Tripathi[1] , Dr. Meghana Chhabra[2] And Dr. RK Pandey[3]
[1]*Senior Manager (IT), Mother Dairy Fruits & Vegetable Pvt. Ltd. , New Delhi, India*
[2]*Associate Professor (School of Management and Commerce), KR Mangalam ,University, Gurugram, Haryana, India.*
[3]*Professor (CS&E) & Dean, School of Engineering & Technology, KR Mangalam University, Gurugram, Haryana, India*
Corresponding Auther: Sarvesh Kumar Tripathi

Abstract:*Challenges with exponential growth of electronic data are being recognized in the world. Such challenges have been observed symmetrical to the challenges of human population growth and physical inventory growth in industrialized world. A policy frame work for selective retention of data has been presented to challenge existing paradigm of "Digitize all, Store all, for ever" . Parallels are drawn from principles of material inventory management for maintaining narrow band of data for perennial cost benefits and availability of data with high quality. An approach based of 25/25 principle is also presented to make immediate beginning.*

Keywords: *Data inventory management, classification of data, selective data retention, shelf life of data, ABC analysis of data, Corporate memory, 25/25 principle, semantic data, Information as Inventory, Data growth, Data Retention policy, Data Quality, Data provenance.*

Date of Submission: 07-06-2018 Date of acceptance: 26-06-2018

I. INTRODUCTION

We are living in the world of data today, data universe is expanding many folds, presenting opportunities as well as challenges of large magnitudes. We are reminded of interesting observations by Garrett Hardin In the context of geometrically growing human populations, in his popular article "Tragedy Of The Commons". He argued that "Most Rapidly growing populations on the earth today are most miserable". "Population problem is consequence of population". "Freedom to breed is intolerable". "Mutual Coercion Mutually Agreed upon". Freedom to breed Mutually Agreed upon". "No technical solution can rescue us from the misery of overpopulation. Freedom to breed will bring ruin to all " [7].

All of above statements made in 1968 for the geometrical growth of human population hold true for data population after 50 years, in 2018. Problem is expected to be bigger since human population grew geometrically [16], while data population is growing exponentially.

Freedom to breed data and store them under present paradigm of digitize every thing, store all, store forever is feared imitation of tragedy of the commons in the context of data population. Two media reports are being reproduced to share world's views on the fast populating data.

"The Guardian reported on 11.12.2017 'Tsunami of data' could consume one fifth of global electricity by 2025'. "The situation is alarming," said Andrae, who works for the Chinese communications technology firm Huawei. "We have a tsunami of data approaching. Everything which can be is being digitalized. It is a perfect storm. 5G [the fifth generation of mobile technology] is coming, IP [internet protocol] traffic is much higher than estimated, and all cars and machines, robots and artificial intelligence are being digitalized, producing huge amounts of data which is stored in data centers" [21].

The Economist reported under title "Data, Data every where" on 25 Feb 2010.

"But big data can have far more serious consequences than that. During the recent financial crisis it became clear that banks and rating agencies had been relying on models which, although they required a vast amount of information to be fed in, failed to reflect financial risk in the real world. This was the first crisis to be sparked by big data—and there will be more." [22]

Recent history reveals four aspects about human populations after Hardin's days in 1968. 1- Problem with human population growth was recognized. 2- Technology was, is used for controlling human populations. 3- Technology alone can not succeed to achieve population goals, human involvement is essential. 4- Intervention begins at source.

World is seeing retarded growth rate of human populations in the relevant societies by confronting the problem through recognition of problem, awareness, technology and human intelligence joining hands and working in synchronization.

Similarly in the context of material inventory, uncontrolled production and accumulation of inventory had been seen a big problem in the past for the profitability of the firms. All above four aspects for human population control were worked upon to overcome the problem of inventory through concepts of MRP, MRP II, ERP and SCM. Computer aided technologies of MRP, ERP and SCM along with active involvement of human intelligence could effectively stabilize the problem of material inventory. Time has come to pay due attention to the problem of Data inventory (Inventory of data).

As opportunities of benefits from large amounts of data are recognized to be huge, academia, industries and Governments are all supporting generation, accumulation and analysis of data. There are special programs for funding and research in the field of big data and data analytics. But benefits are expected to fast diminish with uncontrolled growth in data generation and storage as it has been argued in tragedy of the commons. Per capita availability of natural resources come down with rise in population, similarly per capita availability of knowledge derived from data resources is expected to come down with growing pool of data in centralized data stores. Lots of work and hype is on the display in support of benefits from data (big data) but challenges are not finding deserved place in the published world.

This paper is aiming at review of challenges with fast growing data, presents overview of current approaches to deal with such challenges, drawing parallels with inventory management principles. Hinges at the importance of management discipline to take lead to overcome challenges. A policy frame work is presented as a kick start solution to the problems foreseen with huge data.

II. Challenges With Huge Data Sizes:

2.1. Environmental Challenges:

Challenge of consumption of electricity itself is huge. It is forecast that 'Tsunami of data could consume one fifth of global electricity by 2025'. Further analysis of the forecast gives idea in terms of money. Projected electricity generation in the world will be 34,400 TWh (Tera watt hour) in 2025. Fifth of power generated in the world, 6,880 TWh, will exceed the electricity generated in USA in a year [5]. Estimates are based on the world electricity production from 6,287 TWh in the year 1974 to 23,815 TWh in the year 2014 at annual growth rate of 3.4 % [9]. Cost of 6,880 TWh will be approximately INR 68.80 Trillion (approx. 1.0 Trillion USD) at the prevailing rate of INR 10.0 per KWh charged to commercial entity in India. This amounts to almost three times the Indian budget of INR 21.47 Trillion for 2018-19. We are really going to spent lot of money for storage, retrieval and transportation of data. Carbon emissions are consequential side effects.

Dezyre.com have reported in 2013 that "Only 22% of the data produced has semantic value, of which only 5% of the data is actually leveraged for analysis by businesses. EMC study estimates that by 2020, 35% of the data produced will hold semantic value " [4]. If situation remains same, almost 70 % of data will not have any semantic value in 2025 and wastage in electricity itself will be in the tune of INR 38.00 Trillion, valued at current electricity price in India. That means almost 60-80 % data is a waste and maintained only for draining the scares resources of the earth, unnecessary contribution to the carbon emission and huge challenge to the environment.

Such wastage of resources of gigantic size is result of accumulating information demanded or due for demand and consequent demand for data storage space. If we look at the economics of information, world intellect is working hard on the supply side of information storage, while demand side of the information storage is ignored under the digital data philosophy of **"digitize all, store all and store forever".**

2.2. Quality Of Data:

Technological innovations are developing hardware very fast, almost every three year we are finding new processor, data storage media and applications. Computing power is developing in three dimensions speed, size and complexity. Speed is increasing, sizes are reducing and complexities are rising. Data storage sizes which were requiring a room size are packed in the space of match box. But human desire to digitize all, store every thing for ever is giving rise to opportunities for developments in hardware and software very fast, inducing complexities in the formats and structures of data storage. User creating and storing information, inadvertently, does not care for the size of data and its quality. (S)He does it with perceived abundance of storage media.

Quality of stored data is expected to be poor [4] with respect to the intended **purposes of finding golden needle from hay stack for fault detection or legal evidence and converting straw into gold for business intelligence through data analytics.**

Every tool for business analytics on big data needs preparation of the data. Cleaning of data is vital step for the preparation of such data. As much emphasis is not given to the cleaning of data before storage or in-storage, cleaning of data at terminal end poses limitations and makes it costly.

Stephen Kaisler et al. have identified a typical challenge with quality of big data "An emerging challenge for big data users is "quantity vs. quality. As users acquire and have access to more data (quantity),

they often want even more. For some users, the acquisition of data has become an addiction. Perhaps, because they believe that with enough data, they will be able to perfectly explain whatever phenomenon they are interested in" [14].

Availability of concepts, specifications and tools for cleaning of data at source and various stages like before and after storage of data, similar to TQM in manufacturing will improve quality of data and reduce costs on data retrieval.

2.3. Ownership And Security Of Data:

'Information is power' and 'information is asset' are generally heard phrases. These were true till ownership of data was legally separable and identifiable. Paper based documentation were holding power of information and its owner together. Situation has changed with digitization of information, more so after concept of cloud computing. Resent ransomware attack is a classic example of loss of ownership of information in the form of digital data. Digital data is often fine grained and impractical to validate and stamp ownership on each piece of data. Metadata plays vital role in the issues related to ownership and security of data.

Though size of metadata is increasing but its usage in terms of ownership issues is restricted to the forensic levels. How much help is extracted from such metadata is matter of further research. Huge size of data and lack of metadata retrieval tools is effectively diluting the ownership of different portions of accumulated data. Data are generally stripped off from metadata before its use for data analytics. Diluted ownership of data would dilute the power of information in the hands of intended user and increase the power of unintended user of data.

Computing resource rich societies would exploit power of information collected from resource deficit societies. For example, nations having abundance of electricity, data connectivity networks, costly hardware, trained manpower would tend to enjoy power of information over nations still struggling to have even primitive data connectivity or lack of electricity. High costs on data storage and retrieval would practically enforce one way flow of power of information from resource deficit societies to resource rich societies.

2.4. Management Of Data Storage And Its Transport:

Uncontrolled explosion of structured data along with unstructured data is foreseen a management challenge by many scholars. Kaisler et al. have summarized the big data management challenge as "Data and information provenance will become a critical issue. "there is no universally accepted way to store raw data, … reduced data, and … the code and parameter choices that produced the data." Further, they note: We are unaware of any robust, open source, platform independent solution to this problem." As far as we know, this remains true today. To summarize, there is no perfect big data management solution yet. This represents an important gap in the research literature on big data that needs to be filled" [14].

Another problem with data is the responsibility of costs on data warehouses. IT services provider or IT department is expected to be responsible for costs, while they are not aware of the utility of the data generated. It is user who generates data has to be educated about cost implications on ruthless generation of the data. For individual, it could be person and for firms it can be a line function or functional department.

Storing all data for ever seems good but overdose syndrome will sure come into play and lead to destructive extremes as argued by Alex Coman and Boaz Ronen in the context of management tools and techniques[3].

2.5. Distance Between Data And Actual User Of Data Is Increasing:

Data users can be categorized into two broad categories, one those input data and second, those who retrieve or demand the data for consumption directly or indirectly, after processing the data for knowledge. In case of mobile phone user, laptop or desktop user, single person assumes both the roles. Information input by user is converted into knowledge almost instantly on demand thru few applications residing on the same machine. In such cases data input by machine (machine generated data) are almost non-existent.

For enterprises both the users, data supplier and data consumer are separated apart in many ways. Such users are separated thru hardware, software, middleware, knowledge level and skill sets. In case of data managed on cloud would see such separation in terms of geography and culture too. Data suppliers are generally lowly educated in the context of understanding about end use of information demanded. In case of data is generated by machines like digital probes, camera, digitizer etc. , machines supplying data will not have any idea about end use of data.

Such distance between supplying user and consuming user is increasing with addition of data by volume and variety because of plethora of middleware and manipulation of the raw data. As a result neither supplying user nor consuming user is in position to filter or clean data effectively.

2.6. Multiple Copies Of Same Data:

Best practices for the data volume management suggest to maintain at least 03 sets (copies) of same data as back-up to address outage risk with main system. Such practices enforce maintenance of copies of waste data too. Challenge magnifies in terms of cost when we think of data in Exa and Zeta bytes.

III. Lack Of Concept, Policies And Tools To Decide The Relative Importance Of The Data:

Distance between users discussed above keeps firms indecisive about selective retention of data. IBM noted in their white paper "Lack of retention and disposal policies: Unfortunately, the business side of an organization may not provide IT teams with enough clarity on data retention and disposal policies. Most organizations have a "let's keep it all, someone might need it later" mentality for historical data, which prevents them from exploring cost-effective data retention, hold or purge processes" [12].

Christopher Tozzi of syncsoft.com points out segregation of useful data a grey area, in his blog Best Practices in Data Storage (Part 2): How Long Should Data be Stored? That "You also have to determine how much the data is worth. This is also a bit of a gray area, but it should not be impossible to put a value on certain types of data. For example, you should be able to figure, by reviewing your marketing and sales data, how much it costs you on average to obtain data about a customer or potential customer. You can then use this information to decide when the cost of storing the data exceeds the cost of generating it, and plan your data retention policy accordingly" [23].

Hicks has noted in his work that "Waste within the context of information management is less clear and not generally as visible. Therefore, prior to the application of lean principles to the processes associated with information management, it is first necessary to develop an understanding of waste within the context of information management and characterize the types of waste present within the overall information management system and infrastructure "[8].

Kalfus at al. have argued in his work that "Age of data is often ignored in making storage decisions. Older facts take up storage space, impede access to needed information and may mislead understanding a reality" [14].

As argued by Kalfus et al. and IBM, core problem with uncontrolled data growth lies in the fact that firms and individuals do not have principles, policies and tools to effectively get rid of waste from their data warehouses. It will be noteworthy that real utility of the data is best known to the actual user of data and not to the computer scientists.

Had user been equipped with knowledge and tool to deplete the data in real terms, data sizes could have been much less. In order to compare the analogy with data packet on the internet, Had there been no concept like 'time to live' (TTL), internet could have been clogged by now with historical data packets.

Market forces are working in support of maximizing the digital data for perceived competitive advantage over the competitor through data analytics. Consideration for the challenges has taken back seat. This paper is a beginning for finding answers to rational questions below.

Can we store infinite ? Can we store infinite for ever ? At what Retrieval costs? Utility by accident or pre-planned ? Who decides, What to store, for how long ?

IV. Current Approach For Dealing With Huge Data:

Scholars had foreseen the problem in past and proposed few solutions. Most of the solutions are centered around technology and methodology for increasing storage capacity and speed to retrieve data. Such solutions fell short to deal with the problems with rising data size and costs for individual and firms. Present day approaches are elaborated below.

2.7. Add Hardware:

First and foremost solution to the challenges posed by data growth is add hardware. Recent past has shown us that addition of hardware has offered limited solution despite reduction in costs on hardware. Enterprises created data centers initially and slowly abandoned the idea of owning its own data center, instead opted to move to cloud computing and rented storage. Though cost of hardware is reducing, over all IT budgets are increasing.

2.8. Cloud Computing:

In order to deal with risings costs for data handling by single entity, solutions appeared in the form of "Cloud Computing" by 2006. Cloud computing is all about exploiting economy of scale in data handling. Cloud computing is solving problem of volume of data handling upto certain extent but giving rise to more serious problems related to the security and ownership [17] . Costs on data storage, retrieval and connectivity are still rising. Unaccounted environmental costs are not yet under consideration. Moreover, on-cloud computing models will begin imposing limitations as infinite storage is possible at infinite costs.

Jack Rosenblum of Cloudweaks.com has elaborated four pairs of challenges with cloud computing as security and privacy, Interoperability and Portability, Reliability and Availability, Performance and Bandwidth Cost [19].

2.9. **Selective Retention Of Data:**

Since data storage sizes rising exponentially with the premises of "store every thing", solutions based on economy of scale may not last longer. Scholars have appreciated the problem in past and challenged the premise of "store everything" [15]. Kalfus et al. proposed solution to the problem in their work "Selective data retention approach in massive database". Proposed solution draws many parallels between solution for physical inventory and information as inventory. Solutions appreciate concepts of inventory management like WIP , classification of inventory , ABC analysis of inventory, inventory policy to be applied to information / data management of information in process (IIP) e.g. OLTP systems.

Kalfus at al. have recommended a model to contain data size. Their model suggest to selectively retain the data in a database on the basis of predefined profiling of data. Proposed solution has two parts One, profiling of data field (smallest entity of database) and second, policies for retaining the data in the live system based on the relative importance of data segregated thru data profiling. Profile of data determines relative importance of data. Relative importance of the data is suggested to be decided on the parameters like value of information, criticality of information, cost to collect it again. Less important data to be removed from active system either by deletion or by archiving.

Recommendation of Kalfus et al. may be highly beneficial but need elaborate structure for filtration of data for selective retention. Solution has not been seen in practice yet.

2.10. Lean Thinking And Middleware:

There are footmarks in the literature to induce lean thinking in the entire chain of data creation , storage , processing, retrieval to disposal on similar lines of lean manufacturing. Hicks has proposed five principles to develop strategy for lean information management namely value, value stream, flow, pull and continuous improvements [8].

2.11. Classification Of Data:

Data classification is can be defined as the process of organizing data by categories so that it may be used and protected efficiently. The classification process not only makes data easier to place, locate and retrieve – data classification is of particular importance when it comes to risk management, compliance and data security. Time has come to redefine data classification as "process of organizing data by categories so that it may be used, protected and disposed off efficiently.

Data classification involves tagging of data, which makes it easily searchable and trackable. It also eliminates multiple duplications of data, which can reduce storage and backup costs, as well as speed up the search process. Classification of data assumes special importance when data is huge by size and variety.

Old wisdom suggests to break the problem in small pieces if it is very big. Variety of methods for classification, tagging and segregation are currently in practice. Most of them are proprietary solutions from big software vendors. There is scope of separate study to review all these methods and attempt to unify them under uniform nomenclature. Few of them are being discussed here .

2.11.1. CLASSIFICATION BASED ON TEMPERATURE OF DATA:

IBM have proposed a classification of data based on its temperature. Symbolic temperature of the data is determined on the basis of the frequency with which particular class of data is accessed or retrieved. Data in temperature based classification has been classified in five classes Hot, Warm, Cold, Cooler and coolest as shown in figure -01 [12].

Proprietary solution from IBM offers to deal with huge data by classifying and segregating the data based on its temperature and storing it in cost effective data storage mediums. Hot data is proposed to be stored in OLTP while other data can be maintained in less costly mediums. But there still exists problem of overall size of data.

2.11.2. LOCATION BASED CLASSIFICATION OF DATA:

IDC classifies data based on the location of data creation.

"Core Data refers to designated computing data centers in the enterprise and cloud. This includes all varieties of cloud computing, including public, private, and hybrid cloud. It also includes operational control centers, such as those running the electric grid or telephone networks.

Edge Data refers to enterprise-hardened computers/appliances that are not in core data centers. This includes server rooms, servers in the field, and smaller data centers located regionally for faster response times.

Endpoint refers to all devices on the edge of the network, including PCs, phones, cameras, connected cars, wearables, and sensors" [11] .

Figure -02 shows the projection of relative size of data till 2025.

IDC's data classification is good for presenting statistics of data growth and an idea for the priority for the efforts to control the data. IDC suggests to clean and filter the end point data which will have cascading effect on the edge and core data in terms of its reduced size.

2.11.3. CLASSIFICATION OF DATA BASED ON SECURITY NEEDS:

ISO 27001 suggests classification of information namely Classified, Restricted, Internal and Public. Such classification is intended primarily to serve the purpose of information security but does not attempt to contain the growth of data [9].

In most cases, the asset owner is responsible for classifying the information. This is usually done based on the results of the risk assessment. Higher the value of information, higher is the consequence of breaching the confidentiality. Data with highest risk should be classified in higher class e.g. classified. Data with lower risk to be classified in lower class e.g. public. Framework for the classification of data for the purpose of controlling data growth has yet to evolve . Parallels with physical inventory can be drawn for such evolution.

2.11.4. METADATA AND TAGGING OF DATA:

Metadata (Data about data) for the digital information is also growing along with the growth of data. There are standard formats for the metadata of data files created through particular application. All applications have their own formats and content of the meta data. Most of the metadata is generated automatically by the application. There is hardly any intervention of the user adding the information.

Role of metadata for handling big data and analytics is being recognized widely by technologists working on big data. Apart from technical aspects metadata has vital usage for investigation and settling legal issues. Importance of metadata has been comprehensively described by W. Lawrence Wescott II in his work "the increasing importance of metadata in electronic discovery" [25].

Availability of tools for mass extraction of metadata is scares. Same is the case with the options available to the user for appending metadata to the digital data for segregation like shelf life of data. Work seems to be in progress, as following patents for automatic tagging of data were found while literature review.

Patent for Method for automatic tagging of image data Patent No US 20100076976.

Method, apparatus, and computer storage medium for automatically adding tags to document US 20150019951 A1.

IDC has noted "Data tagging, especially automated tagging, is an important aspect of using cognitive systems. Tagging, after all, applies identifiers to information to make it easy to sort, analyze, put in context, and create value. However, data tagging is in its early stages and needs industry standards, additional investment, better industry know-how, and more data scientists on the job. Although not all data would be valued even if tagged, there still exists (and will continue to exist) a large gap between the actual amount of tagged data and the amount that could benefit from tagging" [11].

2.11.5. **INFORMATION AS INVENTORY:**

Evolutions and inventions in the information technology (IT) management, a separate stream of management and inventory management had been intertwined with each other, feeding problem and solutions to each other. Interdependence on each can be seen through time lines of evolutions in the inventory management and information technology management.

Such interdependent evolutions brought in the opportunities and challenges in the field of data management. Cursory view on the size of data can be taken from data bank with Amazon company. "In 2015, Amazon company became largest data store with the most number of servers, 1,000,000,000 gigabytes of big data produced by Amazon from its 152 million customers is stored on more than 1,400,000 servers in various data centers" [4].

World today speaks of big, big and only big data by size. Situation in the area of huge sizes of data is challenging the core function of the management i.e. "Control". Computing infrastructure , which came as aid to help managers in enterprises to cope with inventories, now losing control over data inventory or giving false promise of control over data growth for vested interest.

As inventories are contained through well stablished principles of the inventory management, comparative overview in Table- 1, attempts at observing symmetries or asymmetries between two inventories. Possibility is being explored whether principles of inventory management can come as aid to the data inventory management.

There is unanimity among scholars to retain data selectively. Framework for selective retention is fragmented and hardly in use for containing the data growth. Unified framework for selective retention of data is a possibility by drawing parallels with physical inventory management principles. Attributes on which inventory management principles can be applied are identified in the Table -1 .

All such attributes can be addressed by classification of data for two distinct criteria one, Shelf life of data and second, notional value of the data.

V. Shelf Life Of Data:

Things produced by nature have inbuilt code for its destruction after natural shelf life. Humans have to device codes and protocols for manmade things like Data. People like to retain valuable things for long time if

supported by storage medium, like to dispose off lowly valuable things as soon as possible, if supported by options. Cycle of identification, classification, segregation and disposal of data should be considered from beginning of any project of digitization of information.

All information does not have same value or requirement of retaining it for same duration. For example, greeting messages and pictures exchanged on social media can best have shelf life of 03 days. Few messages, out of all may have special importance, best known to the receiver, may be desired to be retained for life time of user or even after that. Today we have option to retain all for ever or loose all for ever . For the contrary, laborious manual efforts are needed to find golden needle from haystack of data in the personal data store.

A legal rent agreement digital document is less valuable than ownership agreement document, former could be disposed off after 03 years at the most while later is to be retained for ever. Provisions in applications and protocols are needed to give users an option to define shelf life of particular E-mail item, short message, video clip or a file. Dislike in recent past, it is not possible to effectively maintain personal data bank manually due to velocity and variety of data in the form of photographs, video clips, file , mails are being received. Classification of such data has become inevitable. Randomly selected substandard classification framework is creating problems than solving them.

Data at firm level are retained from legal and business point of views. Laws of the country provide provisions of information retention time 08 to 10 year. But laws are ambiguous about classification of data to be retained. Such ambiguity about data retention is interpreted by firms for maximum safety and all data are retained for even longer periods than prescribed by law.

From business point of view, relevant business head seeks retention of all the data for ever. In the absence of costing framework for the data retained for different class of data, business heads tend to maximize data from all classes. Estimated time for data retention for classes of data at firm level is presented in the Table – 02. List is not exhaustive, other classes can be added by the firm.

Time frames of each class of data are to be translated into laws after debate, policies for shelf life of class of data of firms, similar to the inventory policies of the firm. Implementation of such policies aimed at controlling data growth will not be possible manually , software applications and databases protocols will be needed in support. A new approach with automatic identification, classification and disposal of digital waste is needed for managing digital universe. Users need to be given option to define shelf life of each piece of the data.

VI. Role Of Management In Data Inventory:

Before moving to new approach, let us recall principles of inventory management and analogies with the data storage management. Data storage management can be termed as Data inventory management. Symbolic beginning has already been made with the term "data warehousing" which is borrowed from the discipline of inventory management.

As per Oxford handbook of management "Management is more art than a science, and effective managing happens where art , science and craft meet. It involves becoming aware, attending to, sorting out, and prioritizing inherently messy, fluxing, chaotic world of competing demands that are placed on manager's attention" [26].

Management can be understood quickly by core functions of the management planning, organizing, commanding, coordinating and controlling as commonly known as Fayol's 5 functions of management [24]. All these functions are equally relevant to Operation management (OM) and Inventory management branches of the management. There is need to extend them to the field of data inventory management.

Core function of "control"" assumed special interest in early researches for the business organization. In order to have larger control on costs for profitability and competitiveness under competing forces, organizations were subjected to stringent controls in inventory management supported by aggressive research in the area of inventory management. As a result, a sustainable and fully developed body of knowledge about inventory management evolved from primitive concepts of inventory to lean inventory through material requirement planning (MRP), Just In time (JIT) and lean manufacturing.

Control function of management is focus of this work for the uncontrolled growth of data stock in the hands of a firm and individuals. Core function of inventory management is to control inventory costs.

Computer science always came forward to rescue the management disciplines. Now computer science is facing challenges of data tsunami, management discipline should take lead to contain the data tsunami.

VII. Proposed Approach For Containing The Growth Of Data:

Ronen and Spiegler have observed many symmetries between physical inventory and inventory of information in his work "Information as Inventory"[18] . As concept of inventory management is central to the faculty of management, opportunity has been observed to exploit benefits of concepts of inventory management in the problem area of data inventory management.

By the time world aligns itself for elaborate filtrating of data, introduction of few concepts like classification, segregation and regular depletion of data approach similar to that in inventory management is being proposed. There seems unidirectional inertia in piling up volumes of data, most of them have no value. Inertia needs to be broken. Single and foremost priority is to sensitize the organizations for the need of evolving data management policies on the lines similar to the inventory management. Premise of "digitize all, store all and store forever" needs to be challenged through beginning of a policy frame work at the individual and at the firm level. Let technical solutions come in but discipline of management should not remain mute spectator and keep watching the high cost retrieval of meaningful data from the mountains of data garbage.

Objective of containing data growth, specially data without utility can neither be met by machine nor by humans alone. Adoption of manual approach such as "Electronic records retention: Fourteen basic principles" [20] could not gain popularity, probably due to sheer size and complexity involved in data retention. We believe, if people and firms are equipped with policies and tools to take the regular data haircut and regularly eliminate data without utility at data source itself, a linear growth in the data can be achieved with flatter slop. Evolution of data policies could be taken up by the discipline of operation management (OM) with focus on data inventory costs and development of tools can best be taken up by discipline of computer science / information technology. Data growth is not the problem of information technology alone, it is a multi disciplinary management problem.

A policy frame work is being presented for data inventory management which can guide individuals and firms for keeping their data size within control.

Proposed Policy Framework:

- Shift from paradigm of "store all, for ever" to "lean and clean data".
- Develop personal and corporate memory policy with consideration of benefits over costs.
- Firms to have updated info. on data handling costs for each GB data stored in premises or on cloud.
- Costs on data management should be distributed on different lines of business/ functional departments in the ratio of data generated or stored by that function of business.
- Each distinct unit of data like file, data field or data base table row should be tagged with its useful life as Shelf life of data.
- Data should be classified based on its notional value on similar lines to ABC classification of inventory.
- Data management systems to have software tools to delete data whose useful life has been achieved.
- Data management systems to have tools for review and update metadata for their shelf lives and notional value.
- Each business line function to have minimum and maximum data size planning with consideration of the legal compliance and business objectives.
- Data which can easily acquired again should not be retained.
- File system protocols for defining shelf life and notional value of the data set (class of data) are needed.
- Software tools to be developed to find duplicate information in the firms's memory. Some thing like tools which check plagiarism.

VIII. Where To Begin:

As there is near absence of credible research and supporting data in the area of policies for data retention for individual or a firm, implementation of proposed policy framework will be a problem in absence of tested use cases. Beginning is seen as a challenge. In order to overcome start up challenge, 25/25 principle, postulated by Alex Coman and Boaz Ronen [3] can be used.

Each individual and department of the firm should begin with annual target to get rid of 25 % data which does not have any use. 25% data should be identified to be moved from operational system to semi operational system, similar to recycle bin, in a year. Such segregated 25 % data should be removed from semi operational and moved to archive system or deleted in the subsequent year. Achieves to be catalogued with date of creation and shelf life of class of data. Such data identified for deletion or archiving should be reviewed and audited at least at 03 levels in the firm. Software tools for such tasks will be inevitable.

Targets for the data deletion can follow a firm specific strategy or the techniques of data cleaning one such techniques is suggested by Naveen Joshi in his work 4 ways to improve your data quality [13]. Idea is to keep data clean from generation stage instead of cleaning them at terminal end, for analytics. Most Of data is created at edge, it should be filtered at edge itself [11].

Following a learning curve, firms will be able to find an optimum band of operational data after few iterations. In the process, data inventory management will mature with its own principles similar to the physical inventory management. World data bank would see an aggregate effect and retarded growth in the digital data. Typical model for data inventory management, after implementing 25/25 principle, will look like in the graph shown in figure - 03 at a firm level.

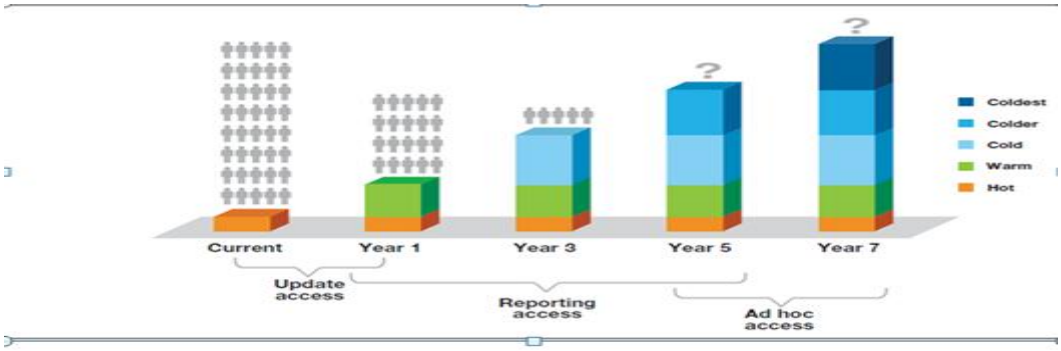
Comparison model shown in figure – 03 considers 30% compounded annual growth in the operational data and annual removal of 25 % of valueless data by archiving or deletion as per data inventory policy of the firm.

IX. Conclusion:

As tsunamis are characterized by catastrophic effects on human societies, tsunami of data too is expected to have similar impacts in data universe. But tsunami of data would be manmade, can be contained by humans. Explosive growth of data will force discipline in data retention, sooner not later. Published literature so far, has very little to offer in terms of principles, tools and policies for containing uncontrolled growth of data. Authors have seen opportunity for dealing problems with huge sizes of data by using principles of material inventory. A data management policy framework has been proposed in this paper, along with a kick start option for data growth reduction using 25/25 principle. Regular depletion of 25 % data at source will help firms determine their cost effective band of data. Aggregate effects will soon be visible in the digital universe. Once firms begin to have data inventory policy and implement it, very high proportion of semantic data in the world's data bank is foreseen. There is possibility of change in the proportion of semantic data of from current 22% to 78% in few years. This will result in cascading effect on the savings in resources like data retrieval, data transmission and data analysis, other than electricity alone. As costs on data handling are proportional to the size of data, cost model on the data handling costs will be same as in figure – 3. Continuous depletion of data will see a data tsunami tamed by 2025 and there after. Future researches with actual data will reveal the effectiveness of the proposed policy framework.

X. Figures and Tables:

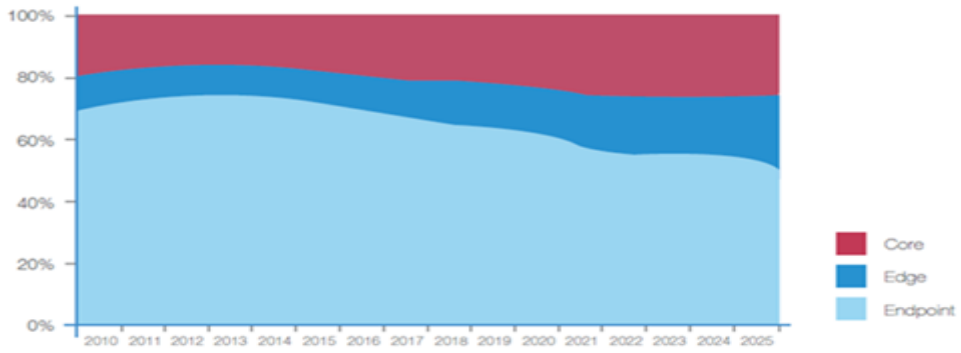
FIGURE – 01: TEMPERATURE BASED CLASSIFICATION OF DATA:



Source: *Benefits of data archiving in data warehouses, Software white paper, IBM, 2013.*

Figure - 01

FIGURE – 02: LOCATION BASED CLASSIFICATION OF DATA



Source: *IDC's Data Age2025 study, April' 2017.*

(Figure -02)

FIGURE – 03: MODEL FOR DATA INVENTORY MANAGEMENT:

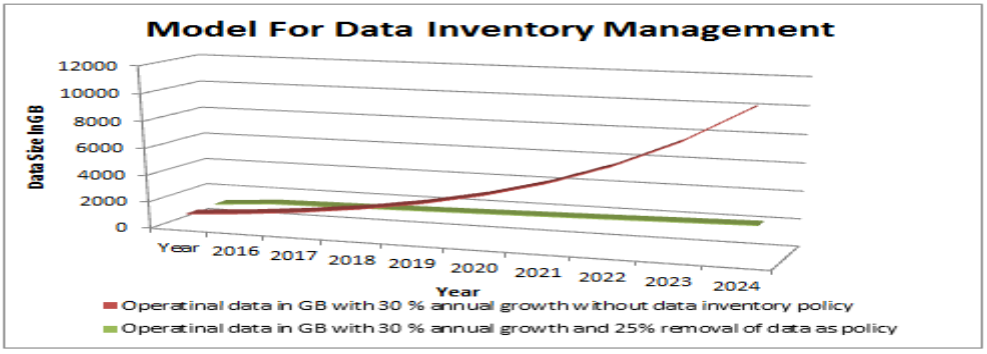


Figure –03

TABLE – 01: COMPARATIVE OVERVIEW OF MATERIAL INVENTORY MANAGEMENT AND DATA INVENTORY MANAGEMENT.

Sr No.	Attribute	Material Inventory management	Data Inventory management	Applicability of inventory management principles
01	Capital costs	Intense	Intense	Yes
02	Variable costs	Intense	Intense	Yes
03	Built-up	Discrete (through orders , receipts and deliveries)	Continuous (through Save and copy commands)	
04	Depletion	Continuous, Through consumption, sale or scrap	Rare, through archiving or deletion	Yes
05	Stock levels	Planned, Defined for firms	Vaguely planned and defined	Yes
06	Maximum stock	Maintained at material level	Maintained at aggregate level	Yes
07	Safety stock	Maintained at material level	Three times of the operational data in the form of back-up and disaster recovery set up.	No
08	Reorder level	Maintained at material level	No relevance	No
09	Lead time	Defined at material level	Defined through SLA's.	
10	Average Working stock	Half of order quantity	Concept not in practice	Yes
11	Shelf life	Critical for batch control materials	**Concept does not exists. There is potential to introduce the concept**.	Yes
12	LIFO and FIFO	Critical for shelf life controlled stocks	Concept does not exists	No
13	Growth	Controlled and linear. Varies in a band of minimum and maximum stocks	Uncontrolled, exponential.	Yes
14	Placement and retrieval costs	Preplanned through location and capacity planning.	**Considerable due to connectivity and band width costs.**	Yes
15	Safety and security	Well defined and documented	Loose in nature due to involvement of external and distributed resources	Yes
16	Ownership	Well defined and documented	Loose in nature due to involvement of external and distributed resources	Yes
17	Duplicity	No Duplicity	Duplicate sets of same data exists in the form of back-up, DR and archives. Subset of same data with employees of the firm also.	Yes
18	Classification criteria for cost optimization	ABC, VEM, Dry and wet, cold and ambient etc. at material or group of materials.	Live on OLTP, archived in cold form	Yes
19	Nature of space occupancy	Vacates space on depletion or partial withdrawal. Reusable.	Partial withdrawal will create copy of same data and occupy space in organizational memory in some form.	Yes
20	Retention philosophy	Minimum stock , JIT, Lean , zero if possible	**Maximum data, Store every thing for ever**	Yes
21	Classical problems	How much to order When to order	What to retain How long to retain	Yes

Table -1

TABLE – 02 : ESTIMATED TIME FRAME FOR DATA RETENTION:

Class of data	Category Of Data	Retention time On OLTP System	Retention Time On OLAP System	Archived Data
		(In Years)	(In Years)	(In Years)
Company Ownership	Registration details, Board Resolutions, Amalgamation / Acquisition agreements, Company Secretory returns, Asset registers,	8	0	For ever
	Rewards & accolades	25	0	For ever
	Policy Documents	5	0	For ever
Legal	Notifications, Court proceedings Judgments	25	0	For ever
Financial	Financial statements, Vouchers	3	5	25
Taxation	Income Tax returns, Sales Tax retunes, Correspondence	3	5	25
Marketing	Brand details, Designs & art work, Recipes, BOMS, General	10	25	For ever
Administration	Personal data, Provident fund data, tax Returns, General	5	5	For ever
Purchase	Quotations and Tenders	3	5	25
	Agreements	5	5	For ever
	Purchase Requisition	1	3	3
	Purchase Order	3	5	25
	Goods Receipt Vouchers	3	5	For ever
	Specifications, Sources Vendors, Market Intelligence	10	10	For ever
Sales	Agreements	10	10	For ever
	Inquiries, Tenders, Sale orders Goods Issue Vouchers, Sales Invoices, Gate Pass	3	8	25
	Customer Data, Distribution data	10	10	10
IT and cross function	Agreements, Licences	5	5	For ever
	Temporary Logs	1	2	0
	E-Mails	3	0	25
	E-Chats	1	0	25
	Complaints	3	5	0
	Files on File server	5	5	25
	Audio , video collection	3	8	For ever
	Market Intelligence	10	10	For ever
	Interphase data of no evidence value	02	0	0

Table -02

REFERENCES

[1]. Businessdictionary.com, Four functions of management, http://www.businessdictionary.com/definition/four-functions-of-management.html, (22.02.2018).

[2]. Chen, D. and Zhao, H. Data security and privacy protection issues in cloud computing. International Conference on Computer Science and Electronics Engineering, Hangzhou, (2012), 647-651.

[3]. Coman, A. and Ronen, B., Overdosed management: How excess of excellence begets failure, Human Systems Management, 28, (2009), 93–99.

[4]. Dezyre.com, Big data timeline- series of big data evolution, https://www.dezyre.com/article/big-data-timeline-series-of-big-data-evolution/160 , (17.02. 2018).

[5]. Enerdata (2017), Global Energy Statistical Yearbook 2017, https://yearbook.enerdata.net/electricity/world-electricity-production-statistics.html (02.03.2018).

[6]. Gantz John and Reinsel David, The digital universe in 2020: big data, bigger digital shadows, and biggest growth in the far east, https://www.emc.com/collateral/analyst-reports/idc-the-digital-universe-in-2020.pdf , (17.02. 2018).

[7]. Hardin Garett. The tragedy of the commons. Science, 162, (1968), 1243-1248.

[8]. Hicks B. Lean information management: Understanding and eliminating waste. International Journal of Information Management , 27, (2007), 233–249.
[9]. ISO, Information classification policy, http://www.iso27001security.com/ISO27k_Model_policy_on_information_classification.pdf , (26.03.2018).
[10]. IEA (2016), Electricity information, https://euagenda.eu/upload/publications/untitled-69168-ea.pdf, (01.02.2018).
[11]. IDC, (2017), Data age 2025: the evolution of data to life-critical, https://www.seagate.com/files/www-content/our-story/trends/files/Seagate-WP-DataAge2025-March-2017.pdf, (05.03.2018).
[12]. IBM, Benefits of data archiving in data warehouses: White Paper, https://www.google.co.in/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&ved=0ahUKEwiJyquA0onaAhVJPI8KHXRgBqsQFgg2MAA&url=https%3A%2F%2Ftdwi.org%2F~%2Fmedia%2FD1E89D3CCDED45389656406F3413F28E.PDF&usg=AOvVaw1QzKnv-7dPqU8DvofrthhQ, (26.03.2018).
[13]. Joshi Naveen, (2016), 4 ways to improve your data quality , Https://www.allerin.com/blog/4-ways-to-improve-your-data-quality, (27.02.2018).
[14]. Kaisler, S. Frank, A. Alberto, J.E. William, M., Big data: issues and challenges moving forward, https://pdfs.semanticscholar.org/a9a8/4242cf78a9b277fce869f8bbfa7bfacfbc38.pdf, (06.03.2018).
[15]. Kalfus Orly, Ronen Boaz, Spiegler Israel. A selective data retention approach in massive databases. Omega, 32, (2004), 87-95.
[16]. Malthus, T. R., An essay on the principle of population [1798, 1st ed.], http://oll.libertyfund.org/titles/malthus-an-essay-on-the-principle-of-population-1798-1st-ed , (02.01.2018).
[17]. Onyshkevych Sev, Five data center issues; 2013, http://www.datacenterknowledge.com/archives/2013/01/08/top-five-data-center-issues-dcim-to-the-rescue, (05.12.2017).
[18]. Ronen Boaz, Spiegler Israel . SOS Information as inventory, Information & management, 21, (1991), 239 – 247.
[19]. Rosenblum Jack, Top five challenges of cloud computing, https://cloudtweaks.com/2012/08/top-five-challenges-of-cloud-computing/, (28.02.2018).
[20]. Stephens, David O; Wallace, Roderick C. Electronic records retention: Fourteen basic principles. Information Management Journal, 34, 4, (Oct 2000), 38-52.
[21]. The Guardian, Tsunami of data' could consume one fifth of global electricity by 2025, https://www.theguardian.com/environment/2017/dec/11/tsunami-of-data-could-consume-fifth-global-electricity-by-2025, (11.12.2017).
[22]. The Economist, Data, data every where, https://www.economist.com/node/15557443, (06.03.2018).
[23]. Tozzi Christopher, http://blog.syncsort.com/2017/07/big-data/data-retention-data-storage-length/, Best practices in data storage (part 2): how long should data be stored?, (27.02.2018).
[24]. Van Vliet, V., Five functions of management (Fayol), https://www.toolshero.com/management/five-functions-of-management/, (201.03.2018).
[25]. Wescott II W Lawrence, The increasing importance of metadata in electronic discovery, http://jolt.richmond.edu/jolt-archive/v14i3/article10.pdf , (10.03.2018).
[26]. Wilkinson, A. (ed.), Armstrong, S. (ed.), Lounsbury, M. (ed.),. The Oxford Handbook Of Management , New York : NY, Oxford University Press, 2017.
[27]. Coman, A. and Ronen, B., Overdosed management: How excess of excellence begets failure, Human Systems Management, 28, (2009), 93–99.

Authors:

1. Sarvesh Kumar Tripathi corresponding author) has 28 years of diversified technology experience at small to large, Government, Co-operative and private, enterprises; technical and personnel management capability at different levels of the organizational structure; hands-on experience in green field project management, production management, commodity trading, ERP implementation and ERP support; currently pursuing PhD along with work responsibility at senior management position.

2. Dr. Meghana ChhabraAssociate Professor (School of Management and Commerce), KR Mangalam ,University, Gurugram, Haryana, India.;

3. Dr. R. K. Pandey

Professor (CS&E) & Dean, School of Engineering & Technology, KR Mangalam University, Gurugram, Haryana, India ; Dr.Ramkinkarpandey@Gmail.com

Sarvesh Kumar Tripathi "Taming Tsunami of Data by Principles of Inventory Management "IOSR Journal of Business and Management (IOSR-JBM) 20.6 (2018): 01-12