Computer scientists say it's time to start looking at treatment of data waste

July 19 2011, by Bob Yirka

(PhysOrg.com) -- As anyone who has ever used a Windows based computer for any length of time knows, the longer you have it, the slower it goes; this is because of the accumulation of data files and entries in system logs; information that in many cases isn't really necessary. Thus, our computers slow down due to the accumulation of "waste." Now, two computer scientists from Johns Hopkins University have published a paper on *arXiv*, where they argue that data waste management on computer systems could, and should be handled similarly to the way physical-world waste is managed.

In their paper, Ragib Hasan and Randal Burns pick up where computer scientists at Cornel University left off after discovering in 1999 that up to 80% of files written to the hard drive by the Windows NT operating system were deleted within five seconds of being created.

Hasan and Burns analyzed three computers: a MacBook laptop, a desktop running Ubuntu Linux and a Fedora Linux fileserver in the University Library (Linux is a variant of the Unix operating system used primarily at educational and research institutions). Their intent was to find out what percentage of the files on each of the computers had not been accessed since their creation. They found that the percentages for each were: MacBook: 20.6, Desktop: 47.4 and Server: 57.1 and that the percentage of disk space used for each was 98.5, 38.1 and 99.5 respectively; clearly indicating that a large number of files using a lot of disk space had never been used again once being created. This is clearly an inefficient use of resources.

It is for this reason that the duo suggest a new approach be used for data waste, one that takes advantage of the research already done with physical waste; specifically, they suggest a pyramid approach be used, similar to the one put in place by physical waste management companies. At the bottom of the new pyramid would be the worst case scenarios, then moving up, the next best and so on till reaching the top, and that they be labeled as such: Dispose, Recover, Recycle, Reuse and Reduce, with zero data waste being the optimal goal.

In this case, Dispose is just that, erasing the data, Recover refers to extracting usable components, Recycle would be refurbishing component for reuse, and Reuse would be using those recoverable components in another way, and Reduce, the ultimate goal would be creating software that doesn't create waste data in the first place.

Besides slowing computers down due to I/O bottlenecks, data waste can also contribute to faster burnout times for flash technology, which have a limited number of lifetime write/rewrites before dying, something the authors point out, will likely become more important as such technology is increasingly being used in hand-held computing devices.

More information: The Life and Death of Unwanted Bits: Towards Proactive Waste Data Management in Digital Ecosystems, Ragib Hasan, Randal Burns, arXiv:1106.6062v2 [cs.ET] arxiv.org/abs/1106.6062

Abstract
Our everyday data processing activities create massive amounts of data. Like physical waste and trash, unwanted and unused data also pollutes the digital environment by degrading the performance and capacity of storage systems and requiring costly disposal. In this paper, we propose using the lessons from real life waste management in handling waste data. We show the impact of waste data on the performance and operational costs of our computing systems. To allow better waste data

management, we define a waste hierarchy for digital objects and provide insights into how to identify and categorize waste data. Finally, we introduce novel ways of reusing, reducing, and recycling data and software to minimize the impact of data wastage.