Go further, faster®

The "Tech"tonic Shift

Dale Wickizer
Chief Technology Officer
U.S. Public Sector
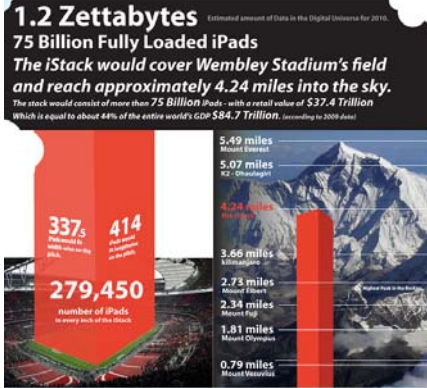
NetApp Confidential — Limited Use

Today I want to talk to you about the "Tech"tonic shift occurring to traditional enterprise applications as well as the IT organizations that manage them.
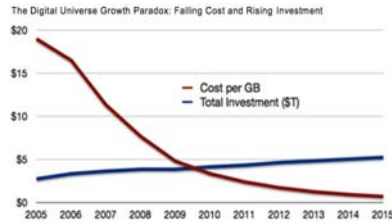
That shift is being driven by an explosion of data being generated and consumed in the world.

Data has grown by a factor of 9 over the past 5 years, crossing 1.2 ZB for the first time! (If anyone wonders what 1.2 ZB is, Wikibon has this great graphic, showing it is the equivalent of 75 billion fully loaded iPads, stacked end-to-end and side-by-side, covering Wembley stadium, in a column more than 4 miles high). This year it will grow to 1.8 ZB.

More than 90% of this data was unstructured and much of machine generated, in response to data stored by end users.

Over the next decade, this data growth is expected to accelerate, increasing by a factor of 50. Over the same time, the number of files is expected to increase by more than a factor of 75, which will break most traditional file systems.

Sadly, the number of IT professionals will only grow by less than 50% and IT budgets will remain relatively flat!
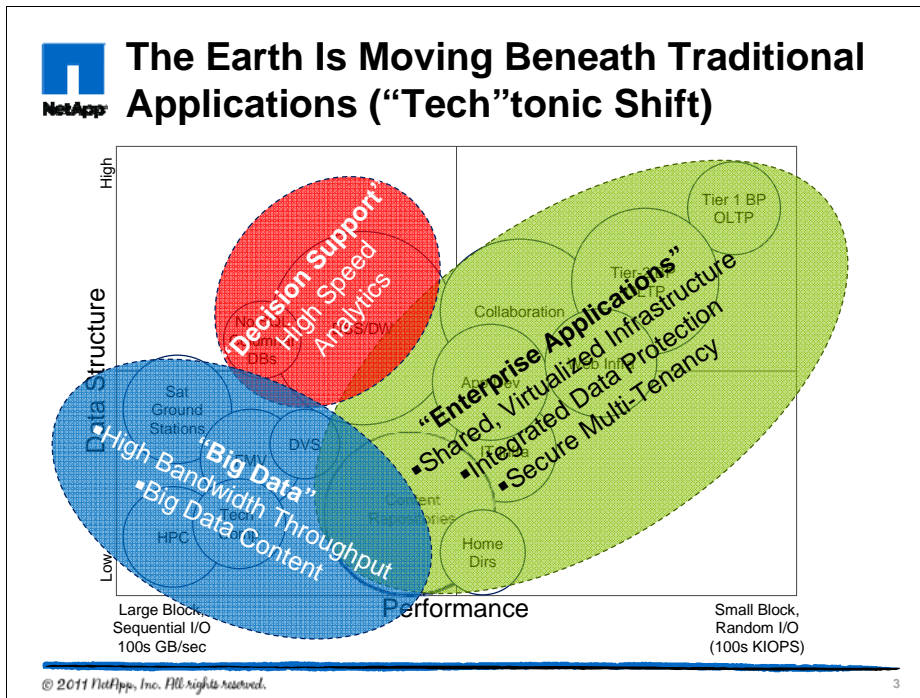
Because organizations don't know what to throw away, they are keeping data perpetually. This is putting a great strain on both enterprise IT organizations and the traditional infrastructure and application vendors that supply them.

In fact, I would argue that the earth has already shifted beneath them; the foundation is cracked.
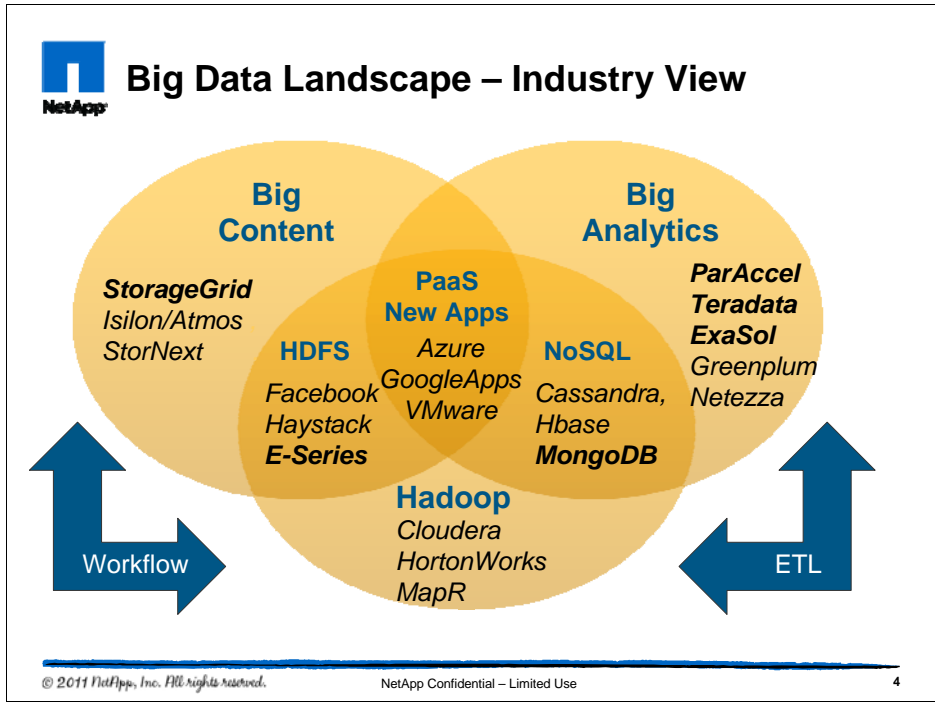
Here is a graph of performance vs. data structure, where I have plotted various types of workloads.

Today's business systems are generally built around highly structured database schemas, built to answer well known business questions very quickly. The infrastructures that support them, particularly the storage, are optimized for small block, random I/O.

With the increasing shift to large amounts of unstructured data to manage, most of it rich media data, the traditional database applications not as effective, for several reasons:

- Large amounts of unstructured data resist being stuffed into structured database applications

- Increasing numbers of unstructured files quickly reach the inode limitations of the file systems running under these applications, requiring them to be split into multiple instances.

- Lastly, rich media requires scale out storage technologies than can handle very high sequential workloads, in addition to small block, random I/O. These scale out storage systems leverage cheap SATA disk drives that are optimized for sequential workloads, along with some type of flash-based caching to handle small block, random read workloads, such as metadata.

And, as the data goes, so goes the metadata (the data about the data). How does one search, index and find answers in all that data, particularly across distributed repositories? The pace at which business questions are changing is accelerating and a lot of the data is more "real time", making it difficult for DBAs to keep up and write schemas that are relevant. Hence we see increased use of high speed analytics.

Big Data Landscape – Industry View

Big Content

Big Analytics

StorageGrid
Isilon/Atmos
StorNext

PaaS
New Apps

ParAccel
Teradata
ExaSol
Greenplum
Netezza

HDFS

Azure
GoogleApps
VMware

NoSQL

Facebook
Haystack
E-Series

Cassandra,
Hbase
MongoDB

Hadoop
Cloudera
HortonWorks
MapR

Workflow

ETL

We already see enterprises beginning to adopt High Speed Analytics and and scale out storage technologies to provide new Decision Support methodologies.

We see new non-traditional NoSQL and columnar database applications coming online everyday, like MongoDB, Cassandra, ParAccel and Hbase. These applications leverage the fast search and index capabilities of Hadoop. They are able to process workloads 100 times faster than Teradata, Netezza or Oracle Exadata over all types of data (structured and unstructured) at a fraction of the cost.

So What?

Impacts of Big Data In Large Enterprises:

- *Increased tendency toward distributed repositories*
  - Network bandwidth will lag behind demand
  - Access data where it "lives" vs. from central repository, saving time, bandwidth and money
- *Metadata Management*
  - Distributed repositories put more strain on the metadata layer
  - Shift away from traditional database / enterprise applications
  - Shift toward NoSQL / MapReduce based metadata solutions
- *Data Management*
  - Shift toward object-based storage with POSIX-compliant & CDMI access, to overcome inode limitations
- *Data Center Challenges*
  - Denser (> 1.8 PB), heavier (> 3000 lb), power hungry (> 15 KW) racks
  - Containerized DCs more efficient cost effective than traditional DCs

5

So What?

Well there are several trends that Big Data will drive in large enterprises:

- Higher bandwidth data will result in more, large distributed repositories as network bandwidth lags behind. Tendency will be access data where it is generated and lives, vs moving it to a central location to save time, money and bandwidth.

- As already mentioned, unstructured data puts a strain on existing enterprise applications. There is already a shift beginning towards disruptive solutions based on NoSQL and MapReduce; distributed repositories put even more strain on the metadata layer. Look for a shakeup in the application vendor space as traditional vendors get kicked to the curb.

- Similarly the glut of more data will put strain on IT organizations, both from a management and an environmentals standpoint. Dense, object-based storage and new containerized data centers are cost-effective and efficient means that can help.

Thank You

Dale Wickizer
Chief Technology Officer
U.S. Public Sector, NetApp, Inc.
Dale.Wickizer@netapp.com
(703) 864-4805

NetApp Confidential — Limited Use 6