PROACT

Proact whitepaper on

Big Data

Summary

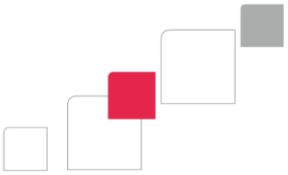
Big Data is not a definite term. Even if it sounds like just another buzz word, it manifests some interesting opportunities for organisations with the skill, resources and need to analyse humungous amounts of data. The challenge is twofold: (1) collect and access the data and (2) analyse the data. Technically, this means:

It is not enough to store data and then manage it – storage, management and availability of data is one unified challenge.

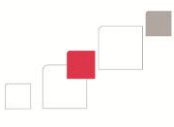
Our current vendors EMC, NetApp and VMware are well positioned in the Big Data marketplace. Proact is no new-comer in the field of Big Data Storage. We have delivered systems for humongous amounts of data with hard-to-meet I/O demands since the 1990-ies.

OPEN INFORMATION

Jakob Fredriksson, Proact IT Group AB, 2013-01-30, version C



Contents



1 What is Big Data?

The term “Big Data” originates from the open source community that tried to develop analytical processes that could incorporate unstructured data and were faster and more scalable than the data warehousing solutions at that time. The term “Big Data” has since developed and as a buzz word it covers a lot more. “Big Data” is nowadays used to refer to a number of advanced data storage, access and analytics technologies aimed at handling high volume and/or fast moving data in a variety of scenarios.

The recent years’ discussions around, and development of, “Big Data” solutions can be divided into two categories:

- Big Data Storage
- Big Data Analytics

1.1 Big Data Storage

Big Data Storage is the storage that supports Big Data Analytics to be performed. Big Data Storage is all about the regular technological discussions that we in Proact are well familiar with:

- I/O performance, latency
- Direct attached storage vs. SAN or NAS
- Data Management on the infrastructure layer and how it interfaces/interacts with the applications

1.2 Big Data Analytics

Big Data Analytics come from web applications but are now entering segments far away from the web companies. It is fair to say that there are opportunities for Big Data Analytics in all major industry segments.

The reason for considering Big Data Analytics is the opportunities to gain business insight from analysing the humongous amount of data that the relational database management systems (RDBMS) cannot handle (more on this in the next section).

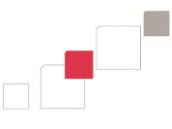
1.2.1 RDBMS vs. Big Data

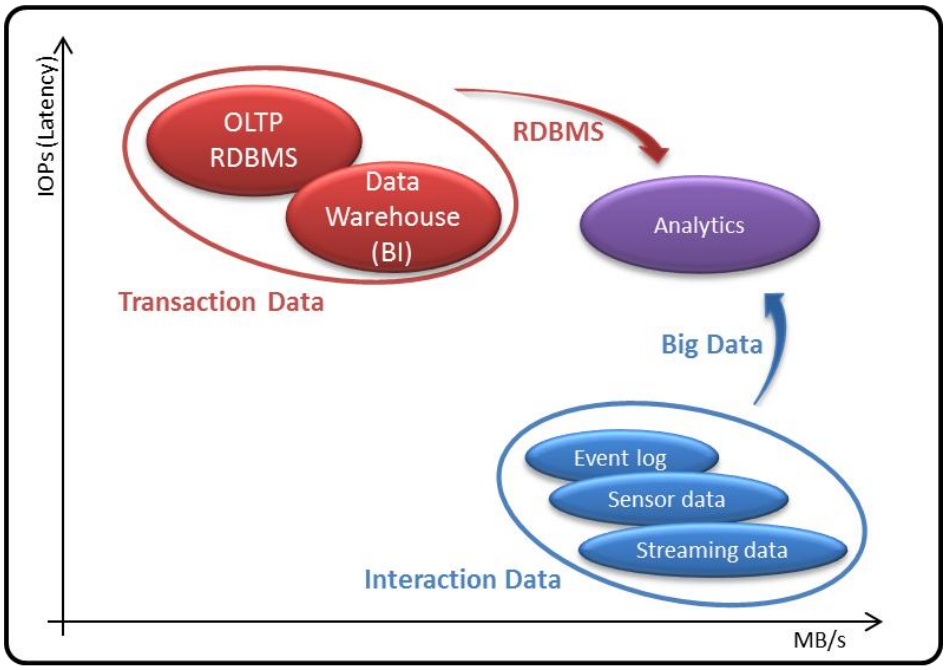
In the previous section RDBMS was discarded as old and insufficient. In this section you will understand why.

It is not enough to state that RDBMS is focused on Online Transaction Processing (OLTP). The real reason for the failure of RDBMS when it comes to the data that is covered by the term “Big Data” is that it is difficult to store, manage and analyse using RDBMS technologies, because RDBMS:

- only handle tabular data
- does not parallelise well enough to accommodate commodity HW clusters
- Does not benefit from the network speed improvements that has been a lot faster than the improvements of seek time of physical storage
- has difficulties to scale-out efficiently – clustering beyond a few nodes is hard
- does not integrate well with non-relational sources of data
- does not handle large enough databases, at least not yet

As you can see from the list above, RDBMS has problems to benefit from the inventions and development in networking and scale-out capabilities as well as having hard time coping with disparate data types and the mere size and growth of data.





Big Data solutions does not compete with RDBMS, they complement each other. However, new means of doing the actual analytics are being developed so you can bring results from both sources together.

1.2.2 Who need Big Data Analytics?

Recent technology trends and the growth of the Internet have generated an immense wave of complex data. As a result, companies that seek a competitive advantage must find effective ways of analyzing new sources of information.

In which industries do they have large enough data and/or wants to gain business insights from its data sources that cannot be done with RDBMS? Who has the need to analyse the new sources of information? The usual suspects are: Web 2.0, Media, Biochemical, Oil & Gas, etc.

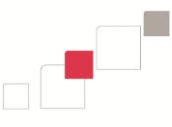
These industries are mainly related to research, but are not the ones that are in focus for the Big Data hype. The focus is instead on the markets that develop a need to improve competitive advantage and responsiveness, for example:

- Telco: Prevent churn[1], by applying social relationships with customers
- Medical: Computer aided diagnostics
- Bank: Risk management
- Retail: Optimization of offerings and discounts.

2 Key technologies that enable Big Data

There are a number of different technologies and solutions that enable the analysis of humongous amounts of data in the market. However, some key technologies stand out:

Technology	How it enables Big Data Analytics
Solid State Disc – SSD	The seek time on hard disk drives is too long.
Scale-out technologies **Full Stack deployments**[2]	Will ensure economics in • processing facilities • storage • implementations, and • management of the systems performing the analytics and store the data. Will also ensure scalability and ability to parallelise computing.
New database technologies (NoSQL, StreamSQL …) **MapReduce/Hadoop**	This is the imperative component for storing, accessing and processing the amounts and types of data that is targeted.



3 Current development segments in Big Data Analytics

This section will explain the four most important areas of development when it comes to Big Data Analytics:

- MapReduce/Hadoop
- Scalable database
- Stream
- Appliance

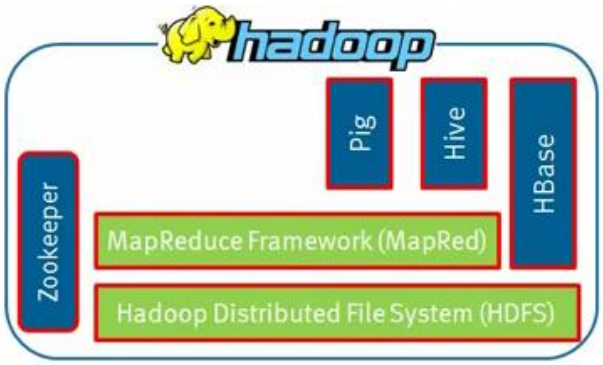
3.1 MapReduce and Hadoop

MapReduce is a parallel programming model (coming from Google) that is suitable for processing of Big Data. MapReduce works as follows, it:

a) splits input data into distributable chunks

b) defines the steps to process those chunks

c) run that process in parallel on the chunks



Thus MapReduce provide the model for breaking down the task of processing large amount of data into pieces that can be distributed into large numbers of compute nodes that work on the jobs in parallel. MapReduce allows scaling through adding more compute notes. This does not imply that all scale-out systems fit in this model since each compute node must have access to all data for this to work.

Note that MapReduce is not an implementation – it is the blueprint. **Apache Hadoop[3] is the most common platform that implements MapReduce**.

Performance optimisation is one important driver for Big Data and Hadoop does it with a different twist: moving the software code to the data and not the data to the software. The design work of Hadoop started off with the idea that it is inefficient to move large amount of data in a cluster, it is much more efficient to deliver the programs to where the data is.

This is made possible by HDFS (Hadoop Distributed File System). Hadoop can with the use of MapReduce and HDFS deliver the programs to where the data is instead of moving the data to the programs.

Hadoop also face challenges. Hadoop deployments that utilize a traditional infrastructure approach, especially direct-attached storage (DAS), often lead to a number of significant issues. Data stored in Hadoop is typically replicated multiple times and distributed across the Hadoop cluster to optimize performance and reliability. Traditional Hadoop deployments run on a cluster of commodity servers for computation (MapReduce), utilizing DAS for data storage (HDFS) and connected together through a network. A single server, referred to as the NameNode, stores all of the metadata for the files stored in HDFS. This traditional approach for Hadoop environments results in a number of challenges for enterprises:

- Storage utilization of DAS is inefficient, a problem made even worse by HDFS replication. In addition, managing large pools of locally attached disks is complex and expensive.
- The Hadoop data staging and loading processes involving DAS is complex and highly inefficient.

- The NameNode is a single point of failure and thereby introduces significant risk in traditional Hadoop deployments. In addition, data backup and disaster recovery processes are often missed in Hadoop environments using DAS.

Additional challenges include:

- technical complexities of Hadoop implementation and management
- inability to independently increase computing or storage capacity
- complex integrations across multiple open source projects

Vendors have of course developed approaches to address these issues (see section 4).

3.2 Scalable database

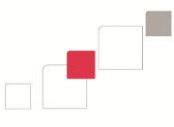
The structured database vendors and communities are not going to lay flat and there is much going on in the structured data communities. Oracle is of course building on their next version and Teradata has acquired a company to add a SQL-MapReduce product to their portfolio. Other initiatives are NoSQL, MongoDB, Terrastor, and SciDB.

3.3 Stream

StreamSQL has been around since 2003 and has the ability to do real time analytics on data streams. It has until now only been used in some niche markets: financial services, surveillance and telecommunications network monitoring. Main vendors in this area are IBM (InfoSphere Streams) and Streambase Systems.

3.4 Appliance

When aiming for the enterprise data centers, we usually see a lot of appliances being pushed from the main vendors. Big Data is no exception. Vendors in this field are: EMC Greenplum, IBM/Netezza, Oracle and Teradata.



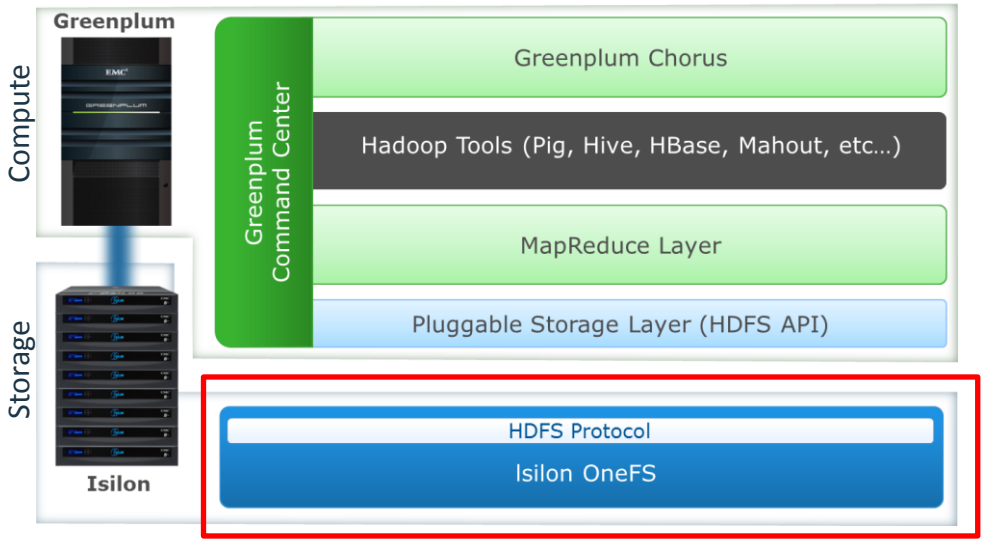
4 Big Data offerings explained

This section is a brief introduction into the Big Data offerings Proact covers.

4.1 EMC: Isilon and Greenplum

EMC have a different approach to solving the issues with traditional Hadoop deployments than most of the competition. EMC avoids the DAS-related issues by avoiding DAS. They can do this since Isilon can integrate natively with the HDFS layer.

This integration enable EMC to combine the scale-out capabilities and resource economics of Isilon and the analytics power of Greenplum HD to create a system that isn't as complex to deploy and manage, that uses storage capacity effective and where computing power and storage capacity can be scaled independently. Greenplum can be offered both as SW only and as an appliance.



EMC has been clever about this architecture since you can use Isilon and its HDFS protocol interface separately integrating any flavour of Hadoop. We expect this to fit most deployments.

4.2 NetApp

NetApp's story[4] is different from other vendors' stories since they use the term "Big Data" in a wider meaning. They do not only focus on the new sources of data and the challenges current analysis solutions have. NetApp also lets Big Data cover the exploding size and management challenges of file systems and databases, bandwidth intense use cases as Full Motion Video, Media Content Management, and Seismic Processing, and HPC systems.

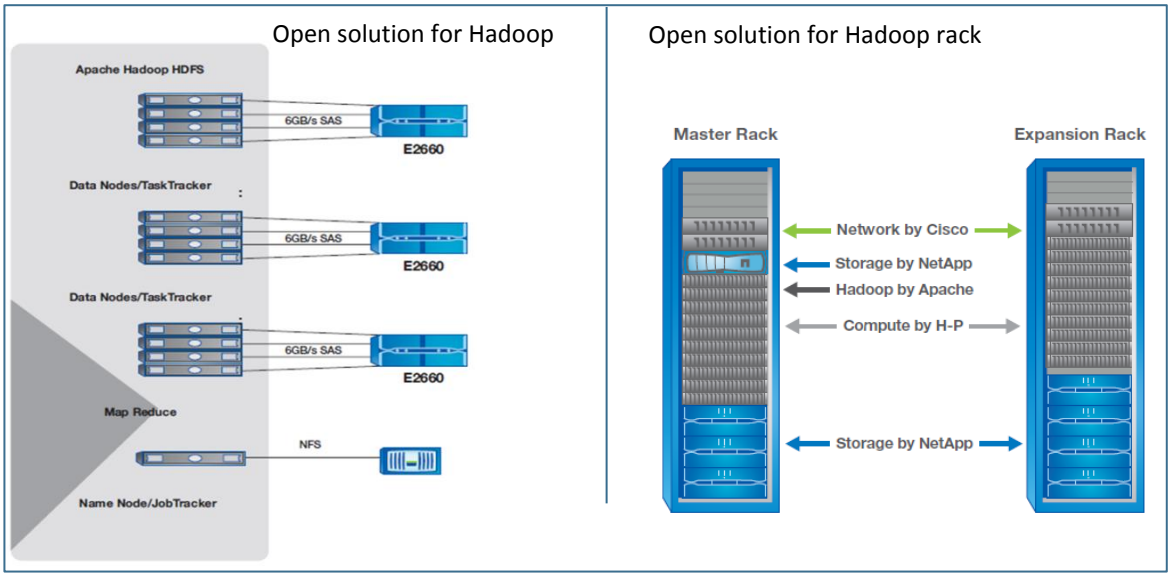
Analytics solutions for Hadoop deployments are delivered with partners: the 2 major ones are Cloudera and Hortonworks, which seems to be the 2 market leaders in this segment.

NetApp-based solutions can be described as follows.

Big Analytics

Providing efficient analytics for extremely large datasets.

Solutions (in partnership with Hortonworks): NetApp Open Solution for Hadoop and NetApp Open Solution for Hadoop Rack.



Big Bandwidth

Obtaining better performance for very fast workloads.

Solution: E-series for storage

Big Content

Delivering boundless, secure, scalable data storage.

Solutions are different depending on challenge:

- File Services: ONTAP 8
- Enterprise content repositories: ONTAP 8 and especially the “Infinite Volume”
- Distributed content repositories: StorageGRID software running on virtualized server infrastructure and E-Series storage systems for storing data

4.3 VMware

In June 2012 VMware launched a new open-source project, called “Serengeti,” that aims to let the Hadoop data-processing platform run on the virtualization leader’s vSphere hypervisor. The reason was explained in the press release[5]:

By decoupling Apache Hadoop nodes from the underlying physical infrastructure, VMware can bring the benefits of cloud infrastructure – rapid deployment, high-availability, optimal resource utilization, elasticity, and secure multi-tenancy – to Hadoop.

However, there are challenges when running Hadoop in a virtual environment[6] and VMware doesn’t pin all their hopes on Serengeti. Serengeti is just one attempt to secure a piece of the Hadoop market. Some other initiatives:

- In February they launched the Spring Hadoop project to help developers write big data applications using Spring Java Framework.
- In April they bought Big Data start-up Cetas Software, which potentially can provide the means to analyse large amounts of data within their virtual or cloud environments.
- In June they presented a reference architecture (along with Hortonworks) for making Hadoop highly available by running the NameNode and JobTracker services on VMs.

5 Turn-key solutions for Hadoop

5.1 Hortonworks[7]

Hortonworks Data Platform (HDP) is a 100% open source data platform based on Apache Hadoop. Hortonworks rely on external vendors for complete.

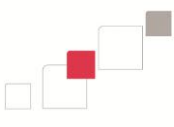
Hortonworks was formed by Yahoo! and Benchmark Capital in June 2011 in order to accelerate the development and adoption of Apache Hadoop.

5.2 Cloudera[8]

Cloudera Enterprise is a turn-key solution for running Hadoop. Cloudera rely on external vendors for complete deployments.

5.3 MapR[9]

MapR distribution for Apache Hadoop is a turn-key solution for running Hadoop. MapR rely on external vendors for complete deployments.



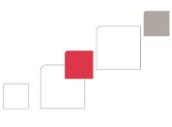
6 Why use Proact's solutions for your Big Data challenges?

Big Data is really not a fair term. The term aims for astronomical humongous amounts of data that is to be ingested, analysed and reported in real time. The challenges when doing Big Data business are many, but I have made the table below as an overview.

Item	Challenge	Solution
Leverage	Ability to understand how the information impacts business in order to transfer it into actions	Model information in current operations with potential strategy impact. Leverage technology to adapt.
Store	"Change" Volume x 2 each 18th month Type: >80% unstructured data Sources of data is disparate	A unified information/content storage methodology that enables management of volume, type and source of information.
Manage	Increased complexity driven by "change"	Tools and services to manage volume, types and sources of information
Analyse	Current solutions are limited to structured data and are too slow	Buy/build real-time analytical solutions for the new sources of information
Use	Multiple access needed to serve diverse use cases	Share, collaborate and act on insights anywhere, anytime and on any device

One important challenge isn't listed above since it is more of a prerequisite for all items in the table but the first one: **availability of data**.

The observant reader has already discovered that which has been Proact's message and strength for some time now: It is not enough to store data and then manage it – storage, management and availability of data is one unified challenge for which there are good solutions!



NOTES

[1] Churn = the rate at which consumers switch service providers for their home TV, phone, Internet, and bundle services. Customers churn habits exhibit patterns of when in their contract life cycle they consider switching providers. In addition, only a handful of motivators — satisfaction with cost, quality of service and customer care — have a major influence on churn.

[2] For example the UCS-based solutions that we sell: FlexPod, vBlock, etc. Can also be called *Converged Infrastructure*

[3] More information can be found at https://hadoop.apache.org/

[4] The story is called *Big Data ABC*, http://www.netapp.com/us/company/leadership/big-data/

[5] http://www.vmware.com/company/news/releases/vmw-serengeti-hadoop-06-13-12.html

[6] Refer to https://wiki.apache.org/hadoop/Virtual%20Hadoop for more information on running Hadoop in a virtual environment and using cloud services for Hadoop.

[7] http://hortonworks.com/

[8] http://www.cloudera.com/

[9] http://www.mapr.com/