Written by : Sunita Sharma
Title : Solution Architect
Dated : April 2018

How Object Storage can improve Hadoop Performance.

This article highlights the importance of Object storage and how it can improve performance with Hadoop.

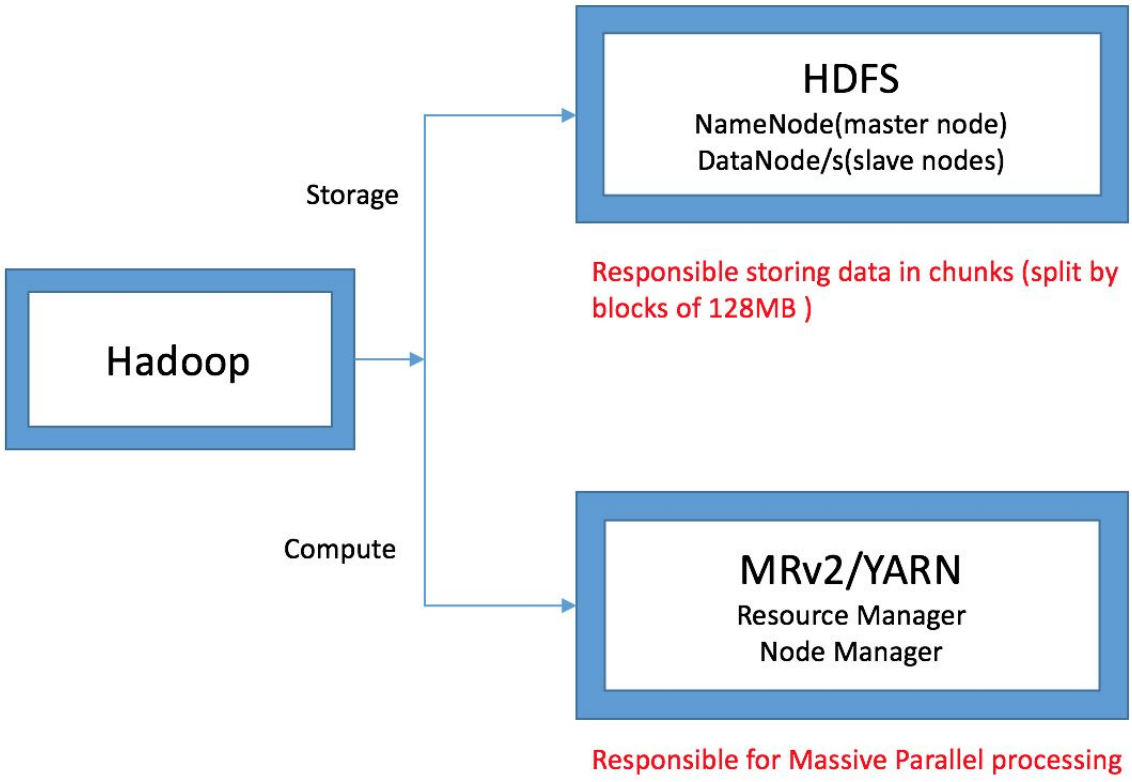
Introduction to Hadoop:

Hadoop is an open source software platform.It is widely used for processing large volume of data processing at massive scale. Hadoop architecture has 2 core components viz. Storage known as Hadoop Distributed File System(HDFS) and compute known as MapReduce(MR).

HDFS has one NameNode and series of DataNode/s. Application data is stored on DataNodes and file system metadata is stored on NameNode.HDFS replicates the file content on multiple Data Nodes based on the replication factor to ensure reliability of data.

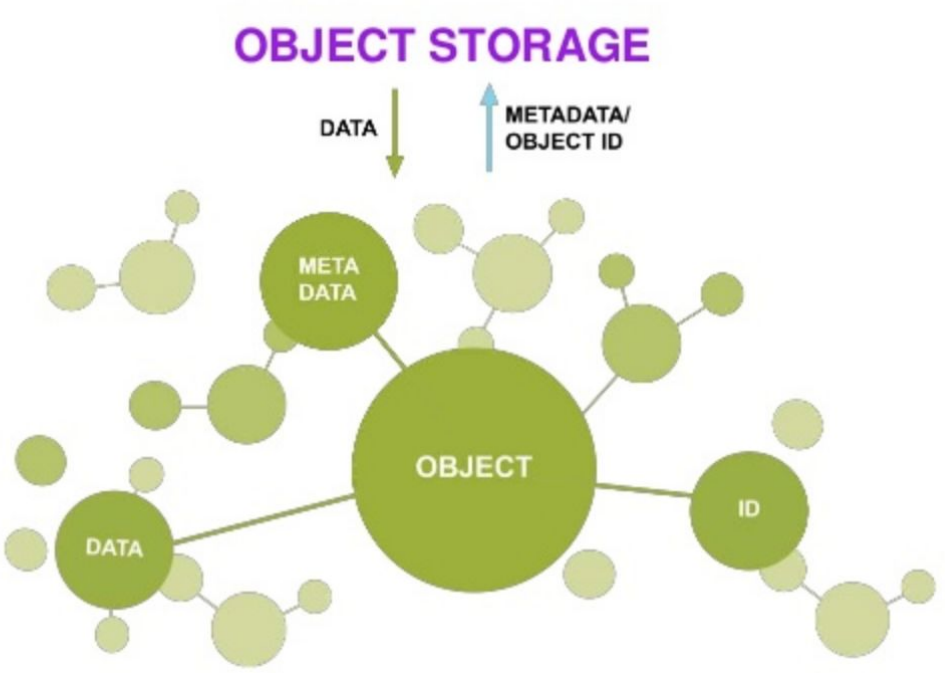
MapReduce is a java-based programming paradigm for distributed processing. Map function transforms the piece of data into key-value pairs and then the keys are sorted with a Reduce function. It is applied to merge the values based on the key into a single output.

The key benefit of Hadoop is that, it brings power of distributed data processing where data is stored. It allows massive parallel processing in a distributed manner.



Introduction to Object Storage

Object storage is a computer data storage architecture that manages data as objects, alternative to file storage which manage data as a file hierarchy, and block storage which manages data as blocks within sectors and tracks. Object storage contains extended metadata. Each object has a unique identifier that lets a server retrieve it from any physical location. Object storage use cases include cloud storage, photos, video, audio and large image files.



Courtesy: Slideshare

The key benefit of Object storage is that, it is low cost, long-term data storage for unstructured data that companies need to keep it for compliance reason, need for long term storage or archive the static data such as photos,videos or images. Each object is stored with metadata and object id that makes it easy to retrieve. It provides simple web services interfaces for access with APIs or http/https.

Current Challenges with Hadoop and How Object Storage can address it:

Hadoop is ever changing and evolving since its origin. The cons is ,continuous evolvement does not make it very stable. While pro is, changes are also bringing new paradigms for compute and storage options.

Some key challenges are :

Scalability:

HDFS do not allow independent scaling.In Hadoop, compute power and storage capacity need to scale in lockstep, meaning you can't add one resource without the other. Object storage on the other hand can scale out easily beyond Petabytes. Data stored on object storage can be easily accessed and processed on Hadoop as needed.

Cost:
A big benefit with object storage is we can separate storage from compute, as a result, larger cluster can be rolled for a smaller period of time to increase throughput, up to allowable physical limits.This separation not only lowers cost but also improves the performance.Object Storage cost is about 1/5 th that of Hadoop platform.

Accessibility and Durability:
Namenode is single point for failure in Hadoop. If namenode fails, it is difficult to access rest of the cluster. With Object storage, you don't need to worry about data accessibility or data loss. Object storage uses erasure coding that helps to prevent the data loss,alternatively data can be made available on any other instance if one instance of Hadoop fails.

Elasticity:
One of the nicest benefits of object storage is it works on pay as you go model. You are only charged for what you put in, and if you need to put more data in, just dump them there. Under the hood, the cloud provider automatically provisions resources on demand. Object storage is elastic, HDFS is not.

New processing paradigms:
Hadoop has evolving since its origin. Apart from Mapreduce that processes data locally on HDFS, It brought in new in-memory processing called Spark. This eliminates need for storing data into HDFS for processing. Data can be stored at low cost in object storage and can be easily accessed in Hadoop for in-memory for processing.

Other Comparables :

Feature	HDFS	Object Storage
Structure	It is File based structure contains blocks to store the data	Object based storage, It has object id and metadata attached to it
Setup	Hadoop cluster is required to read/write data into HDFS	Standalone. Independent to Hadoop.

Latency	Low	High
Storage management	Hands-on management	Storage SLA is provided by cloud vendor. No need to manage storage

Conclusion:

- We discussed various factors in section above that shows object storage to be more desirable compare to HDFS.
- It is not a replacement of Hadoop.
- It complements Hadoop by providing low cost storage alternatives to make processing more powerful.
- It offers simple web services interfaces for access with APIs or http/https.
- It guarantees that the data will not be lost.
- Listed below are few comparables from cloud market leaders that offer Object Storage solutions.

	AWS	Google
Service name	S3 (Simple Storage Service)	Google Cloud storage
Cost	$0.026 per GB	$0.026 per GB
Availability(SLA)	99.95%	99.95%
Object limits	unlimited	unlimited
Max Object size	5TB	5 TB
Hot	S3 standard	GCS Nearline
Cold(Archival)	Glacier	GCS Coldline