Cloud Filesystems

1. Filesystems:

**1. A self-administered networked filesystem**

+ You host your own filesystem and therefore have more scope for settings  
+ Central management, e.g., easier to make backups  
+ Access can be given to ip-adresses, users, etc. Anyone with login data/access can log in and then has access to all the files available for him

- You occupy one of the servers and therefore less computing power for the application  
- heavy network load reduces the performance  
- Single point of failure. If the system breaks down, no operation can be handled   
- Security: NFS uses RPC which is insecure and therefore the server needs to be properly placed behind a firewall. After that you still shouldn’t blindly trust it

**2. A networked filesystem provided by the cloud provider**

The same advantages and disadvantages as above for the the networked filesystem. The only difference here however is that you could now also use the second application server in the way you desire as the filesystem is now running on the cloud. A disadvantage could be the additional cost for the cloud service. But as the cloud is not that expensive normally its just a minor one.

**3. A self-administered distributed filesystem that is synchronized across your hosts**

+ data redundancy. Should one server break done than there is the second one who can handle the requests  
+ should the servers be located at different locations than the user request can be send to the server closer such that the latency gets decreased

- both servers must function as a filesystem and therefore their other functionality gets decreased  
- both servers need to be up to date. Therefore, more network traffic that reduces the performance of the servers

2. Storage Solutions:

**1. Hadoop HDFS**

Hadoop HDFS is a distributed, scalable and portable filesystem. HDFS is the primary storage system used by Hadoop applications. It provides/ensures high performance access to data across highly scalable Hadoop clusters. HDFS consist of five services, Name node, Secondary Name Node, Job tracker, Data node and Task tracker, whereas Name node and Data node are the major services.   
HDFS contains only one **Name Node** that keeps track about the files, where they are stored, where the replications are stored, and gives access to read, write, create, delete, … operations.   
The **Data Nodes** stores the data in blocks as the name already indicated. The data node sends in regular intervals a message to the name node to convey that it is still alive. Should the data node not send a message for a longer duration, the name node replicates the data of that particular data node onto another data node.   
The **Secondary Name Node** creates snapshots of the data of the name node in order to recreate it if it should break down. This is quite helpful as the name node is the single point of failure of this system as it stores all the metadata and is responsible for all the operations on the filesystem.   
The **Job Tracker** schedules map or reduce jobs to task trackers with awareness of the data location. This ensures that the jobs are being processed on nodes that store the data, if possible, such that network traffic gets reduced.  
The **Task Tracker** gets the task to perform from the job tracker and applies the code/ job on the corresponding file.  
Benefits and liabilities of this Filesystem are:  
- Cost effectiveness: HDFS is license free and can be used without any costs  
- Large dataset: HDFS can store a variety of data of any size  
- Fast Recovery: Should a node break down the system can restart/reassign the node/tasks as the name node always is in knowledge about the system and data is stored replicantly. Should the name node however break down, the secondary name node is able to restart/recreate a new name node

**2. GlusterFS**

GlusterFS is a scale-out network-attached storage file system. It has found applications including cloud computing, streaming media services, and content delivery networks. The GlusterFS documentation requires multiple servers, at least three, as GlusterFS is a distributed file system that needs several computers connected to each other. However, in this context, servers can be any form of physical or emulated hardware. Even virtual machines can be integrated which makes it more flexible. The integrated servers are acting as nodes which are connected to each other into a so-called trusted pool. Their memory is provided in the form of bricks, on which drives/volumes are being formed. These can be mounted and used like a normal drive. Computers accessing a server are called clients. A computer can be server and client at once. One of the features of GlusterFS is scalability. Any number of nodes and bricks can be added at any stage. This means that the size of storage can be adapted whenever needed. Also, GlusterFS ensures fail-security through redundancy. The risk of failure is distributed over several systems that can also be located at different locations.

**3. Amazon S3**

Amazon S3 stands for Amazon Simple Storage Service, which is a cloud-based persistent storage. Simple here refers to the feature set and not the ease of use. The user is able to simply put data in the cloud and retrieve it. He doesn’t need to know anything about how it is stored or where. S3 is also not a remote filesystem. It is more primitive as you don’t store files, you store objects. The objects are being stored in buckets and not in directories. This storage has some important differences to a normal filesystem:  
- objects stored can’t be larger than 5GB  
- buckets are existing in flat namespace shared along all Amazon S3 users  
- buckets can be made public  
- third party tools are needed in order to mount S3 storage

**4. Amazon Glacier**

Amazon Glacier is an online file storage web service that provides storage for data archiving and backup. It is designed for long-term storage of data that is infrequently accessed and for which retrieval latency time of 3 to 5 hours are acceptable. Storage is quite cheap with $0,004 per gigabyte per month. The underlying technology used is unknown and subject to speculations.

**5. Google Cloud Storage**

Google Cloud Storage is a RESTful online storage web service for storing and accessing data. The services combine Google’s cloud performance and scalability with advanced security and sharing capabilities. It is an infrastructure as a Service (IaaS), like Amazon S3. In contrast to google drive, the google cloud storage is more suitable for enterprises. Google Cloud Storage stores objects in projects which are organized into buckets. Bucket names and keys are chosen so that objects are addressable using HTTP.

3. What are the security implications of using Amazon S3?

Amazons’ servers are located at unknown locations and access into it is also highly protected such that it is really, really hard for an outstanding person to get into the facility and make some physical damage to the servers. Only the company itself has access to the data and applications running on the cloud. Whoever wants to access it has to double authorize and therefore can’t easily get access. The data is saved at a secure place without any inference to other applications. The data is like in an own private environment, only the company that owns it can access it. Even the cloud provider itself is not able to access or interact with the data. Whenever a virtual server or data base is shut down the whole memory it used is getting zeroed. That means that it is not possible to get access to the data saved there anymore. If you think that this is still not safe enough you can also encrypt the saved data, such that it is even harder to get access to it. Because of this it should be clear that a server in the cloud is save. As the cloud provider has a huge amount of specialists employed it is also likely that their server is more secure than the server your company can ever setup.