(Q1) What is TSM deduplication?

1. It is an optional TSM feature that removes redundant data from a disk-based TSM storage pool. Reducing the amount of backup* data can reduce the cost of storage associated with backup and may allow more data to be stored on disk for faster access.
2. It is important to consider that deduplication is just one method for data reduction. TSM also uses a progressive incremental backup methodology, which only backs up changed data, and supports client-side compression. Additionally, TSM allows exclusion of individual files from backup operations, which further reduces the data involved in these operations.

   NOTE*: References to backup and backup data also apply to archive and space-managed data (space-managed UNIX data can only be used with server-side deduplication).

(Q2) How effective is TSM deduplication?

1. Deduplication effectiveness is usually measured in terms of the ratio of the amount of data before deduplication to the amount of data after deduplication, called the “deduplication ratio”. It can also be expressed as a percentage of data reduction. However, what is more important than just deduplication is the total data reduction of the backup data, which for TSM can include progressive incremental, deduplication, and optionally, compression.
2. TSM deduplication is as effective as any deduplication technology that is available on the market. Deduplication effectiveness is mostly determined by the type of data that is being backed up, and whether the data is unique or repeated. For example, repeated full backups of the same data results in high deduplication ratios, but backing up only changed data (such as with the progressive incremental methodology) results in a lower deduplication ratio. However, with progressive incremental backups the overall data reduction ratio will remain high. Data that is very unique and not backed up repeatedly will typically not benefit from deduplication.
3. TSM deduplication ratios typically range from **2:1** (50% reduction) to **15:1** (93% reduction), and is data dependent. Lower ratios are associated with backups of unique data, and higher ratios are associated with backups that are repeated, such as repeated full backups of databases or virtual machine images. Mixtures of unique and repeated data will result in ratios within that range. If you are not sure of what type of data you have and how well it will reduce, use 3:1 for planning purposes when comparing with non-deduplicated TSM storage pool occupancy. This ratio corresponds to an overall data reduction ratio of over 15:1 when factoring in the data reduction benefits of progressive incremental backups.

(Q3) Is TSM deduplication free?

1. There is no additional software license cost. However, TSM deduplication will require additional resources (memory, sufficient space and disk performance for the TSM database, and CPU).
2. TSM deduplication requires additional processing, either on the client or the server. This additional processing leverages TSM’s internal database. Therefore, it is important to ensure that there is sufficient disk capacity for the TSM database and it is installed on a disk device that can support the I/O performance requirements.

(Q4) When should I consider using TSM deduplication?

1. Consider using TSM deduplication when the following conditions apply:
   - You plan to use a disk-only backup solution (your primary backup storage pool will remain on disk).
   - Your priority is to reduce the amount of disk storage required for backup data.
   - You have a limited bandwidth connection from clients to the TSM server. In this case client-side deduplication is an appropriate solution.
   - You are considering using TSM node replication (available in TSM 6.3).
   - Your TSM database is properly sized for deduplication and resides on a high performing disk array (see additional FAQs for examples).

(Q5) What is the largest amount of data that can be backed up to a TSM deduplicated storage pool?

1. The practical limits of deduplication for each TSM server instance and a given hardware configuration are based on two main factors: (1) the maximum amount of "source" data to be backed up, and (2) the maximum amount of data that will be backed up each day. "Source" data refers to the original data that is backed up along with all versions and copies of that data.
2. Deduplication scalability limits are established for each TSM server instance. The limits apply regardless of how many deduplicated storage pools are configured on a single TSM server. Although there are no theoretical limits, there are practical limits determined by the maximum database size and amount of daily backup data.
3. The practical maximum values depend upon many factors including the resources available to the TSM server (including CPU, memory, and I/O performance). High-performing systems that include solid-state disks for the TSM database will be able to handle greater capacity and workloads than systems with fewer resources. See Q7 for an example configuration. Greater capacity can be achieved with additional hardware resources. The maximum amount of data backed up to all of the storage pools within a single TSM server instance should be kept under 300TB, since this roughly corresponds to the maximum recommended database size of 4TB. The daily maximum backup data is a more critical limit, and this will be determined by the ability of the hardware resources to contain daily processing of backup data. You should consider limiting the daily backup amount to deduplicated storage pools in a single TSM server instance to 4TB or less per day.
4. Backup capacity with deduplication can be scaled out by adding TSM server instances. However, deduplicated data is shared across TSM servers only when TSM node replication or a shared deduplicating appliance is used.

(Q6) How does TSM deduplication affect backup and restore performance?

1. With client-side deduplication, client backup elapsed times can be longer compared to backups to a disk storage pool that is not deduplicated. However when the backup network is constrained, backup elapsed times can be faster when using client-side deduplication. The use of server-side deduplication does not directly affect backup throughput.
2. Throughput for storage pool backup operations from a deduplicated storage pool to a storage pool that is not deduplicated is slower when compared to backing up a storage pool that is not deduplicated.

3. Restore throughput from a deduplicated storage pool is generally slower when compared to restore from a disk based storage pool that is not deduplicated. However, when compared to restore performance from physical tape, restore from a disk-based deduplicated storage pool can be much faster.

(Q7) What are the hardware prerequisites for using TSM deduplication?

As an example, the following configuration will support 120TB of backup data (source data plus retained versions) and 3TB/day of backup ingest.

- CPU: 8@2.2Ghz processor cores or equivalent
- RAM: 64GB (minimum)
- Database disk used: 1.8TB
- Database disk I/O usage: 6000 IOPS sustained (midrange to high end disk array)

(Q8) How do I decide between using TSM's server-side or client-side deduplication?

1. Here are some circumstances when you should consider using client-side deduplication:
    - You wish to achieve the highest potential data reduction, since client-side deduplication can be combined with compression.
    - You wish to distribute the workload across client systems rather than perform deduplication processing in the TSM server.
    - Bandwidth between the client and server is constrained.
2. Here are some circumstances when you should consider using server-side deduplication:
    - The fastest possible backup time is required to meet service-level agreements.
    - You require the shortest possible window for producing non-deduplicating storage pool copies (such as for shipping offsite).
    - CPU resources on the client host system are inadequate to support the additional processing required by client-side deduplication during scheduled backup processing.

(Q9) How do I decide between TSM deduplication and a deduplication appliance?

1. Here are some circumstances when you should consider using TSM deduplication:
    - You plan to use disk-based storage pools.
    - Based on your total backup data and daily backup amount it is more cost effective to invest in TSM server resources than a deduplicating appliance
2. Here are some circumstances when you should consider using a deduplication appliance:
    - You wish to take advantage of deduplication across multiple TSM servers using the same deduplication appliance.
    - Your backup data consists mostly of very large files (greater than 500GB).

(Q10) Which TSM features and options are incompatible or not supported with deduplication?

1. Client-side encryption is incompatible with TSM deduplication. However TSM deduplication can be used together with SSL (encryption of data in flight) or encryption by the storage device.
2. LAN-free backup is not supported for client-side deduplication. However LAN-free backup can be used with server-side deduplication.
3. Simultaneous write
4. Subfile backup
5. UNIX HSM
6. Client side compression should not be used with *server-side* deduplication (since compressed objects do not deduplicate well). However, client-side compression used in conjunction with *client-side* deduplication can provide an effective means to further reduce storage pool data.

(Q11) How do I estimate the TSM database size when I use deduplication?

1. A detailed explanation of how to estimate the TSM database size when using deduplication is available in the following technote: http://www.ibm.com/support/docview.wss?uid=swg21596944
2. Refer to the following table for a rough estimate of TSM database capacity required when using deduplication. This uses the rule of thumb of 150GB of database capacity for every 10TB of backup data.

Total amount of backup data (TB)	Additional database size required for deduplication (TB)
20	0.3
50	0.75
100	1.50

(Q12) How do I determine how much storage I have saved by using TSM deduplication?

1. The easiest way to determine deduplication storage savings is to use the administrator command `“query stgpool f=d”.` The value of the `“Duplicate data not stored”` field will show the amount of bytes saved and the percentage of savings. Note that this value is not updated until after reclamation processing occurs for server-side deduplication.
2. The best way to determine deduplication results is to run the query script available on the TSM support site: **http://www-304.ibm.com/support/docview.wss?uid=swg21596944.** This script provides a summary of deduplication results as well as pending operations.

(Q13) Isn't my backup at risk when my data references "chunks" of data from other files or hosts, and does not store all of the original data?

1. The additional risk that TSM deduplication presents to data integrity is infinitesimally small. Best practices in data protection, such as making copies of backup data, are standard for mitigation of loss of backup data for any reason.
2. TSM provides a number of checks to ensure data integrity for all backup data, including deduplicated data. For chunks to be considered as duplicates, the chunks must have the exact same 160-bit SHA-1 digest and chunk size. TSM also computes and stores a 128 bit MD5 hash value for the entire file (or object) that is being backed up. The MD5 value is used to ensure that the data has been backed up properly, and upon restore this value is used to verify the integrity of the restored data.
3. TSM offers several different levels of protection. Keeping a secondary copy using storage pool backups is recommended for providing the highest level of protection. This recommendation is made in all cases regardless of whether deduplication is used.

(Q14) What TSM server code levels are recommended when using TSM deduplication?

1. You should install the latest TSM server maintenance for your point release when using TSM Deduplication (ftp://service.boulder.ibm.com/storage/tivoli-storage-management/maintenance/server/).