Document Database Design - Embedding or Referencing?

Embedding vs Referencing

- Typically it is most advantageous to have denormalized data using embedding
- There are scenarios in which referencing or normalizing data is beneficial, especially when data from multiple sources needs to be blended for analysis

Scenarios for Embedding

- Data with a 1:1 or 1:many relationship (where the “many” objects always appear with, or are viewed in the context of their parent documents) are natural candidates for embedding within a single document.
- The concept of data ownership and containment can also be modeled with embedding

Embedding Example

- Using the product data example, product pricing (both current and historical) should be embedded within the product document since it is owned by and contained within that specific product
- If the product is deleted, the pricing becomes irrelevant

Modeling Relationships with Referencing

- Referencing enables data normalization
- Provide more flexibility than embedding
- References are usually implemented by saving the _id field of one document in the related document as a reference
- Performance may degrade

Scenarios for Referencing

- Not all 1:1 and 1:m relationships should be embedded in a single document
- Embedding would not provide sufficient read performance gain
- Where a document is referenced from many different sources
- Represent complex many-to-many relationships
- A document is frequently read, but contains an embedded document that is rarely accessed
- One part of a document is frequently updated and constantly growing in size, while the remainder of the document is relatively static
- The combined document size could exceed MongoDB’s 16MB document limit

How to Relate Data when Referencing

- Use $lookup stage in MongoDB Aggregation Framework.
- $lookup provides JOIN capabilities in MongoDB
- Use $graphLookup to recursively look up a set of documents with a specific defined relationship to a starting document