08/07/21
Thursday

BIG DATA

Ribin Sosa John
S4-RMCA
Roll No: 19

2b) In-Memory Storage Device

An in-memory storage device generally utilizes RAM, the main memory of a computer as its storage medium to provide fast data access. The growing capacity and decreasing cost of RAM. The main advantage is storage of data in memory eliminates the latency of disk I/O and the data transfer time between the main memory and the hard drive.

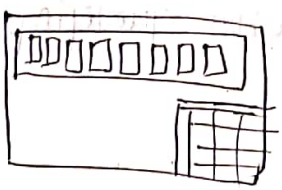
In-memory storage devices can be implemented as:

(i) In-Memory Data Grid (IMDG)
(ii) In-Memory Database (IMDB).

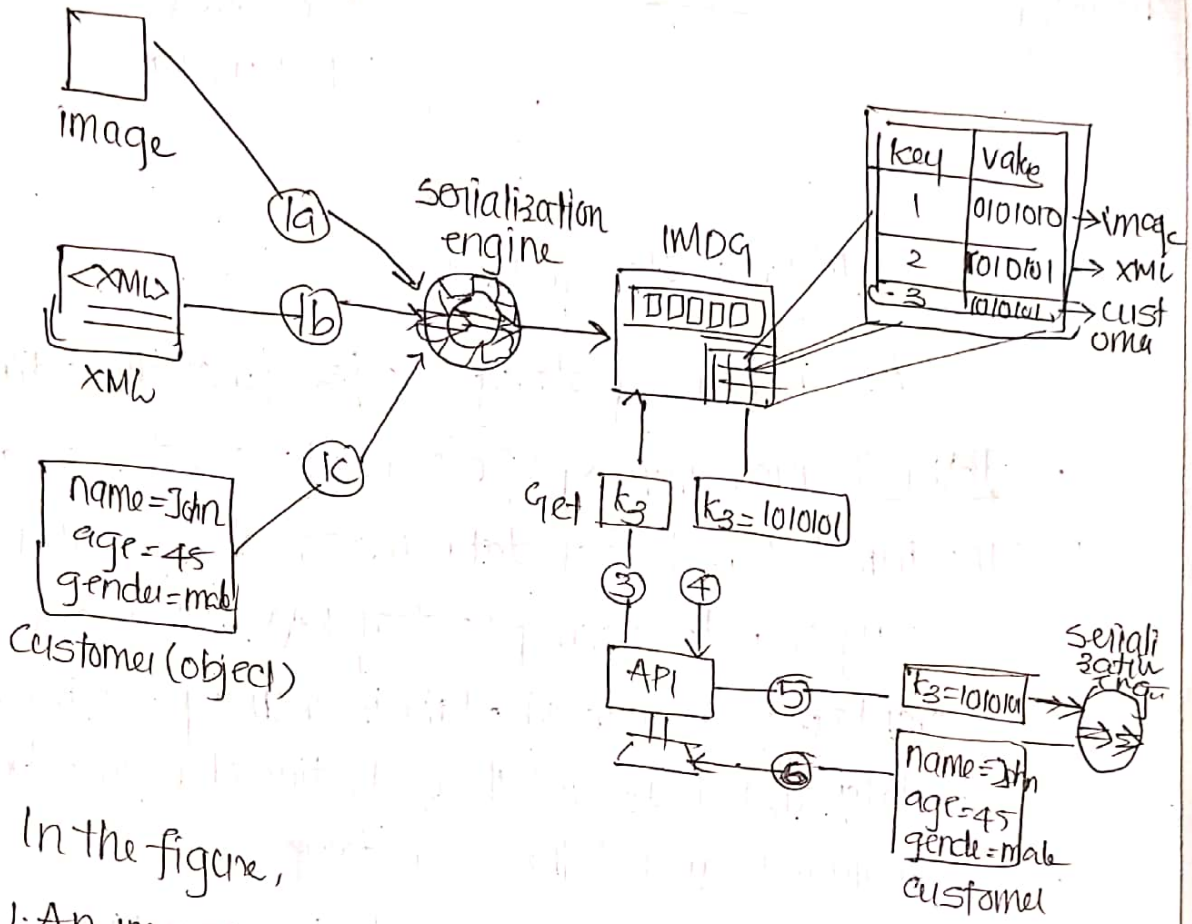
(i) In-Memory Data Grid (IMDG)

It is stored in RAM and the data will be stored in key value pair. This supports schema-less data storage through storage of semi/unstructured data. IMDG provide fast data access. Nodes in IMDG keep themselves synchronized and collectively provide high availability, fault tolerance and consistency. IMDGs scale horizontally by implementing data partitioning and data replication.

~~In the figure:~~
~~An image~~



- IMDG.

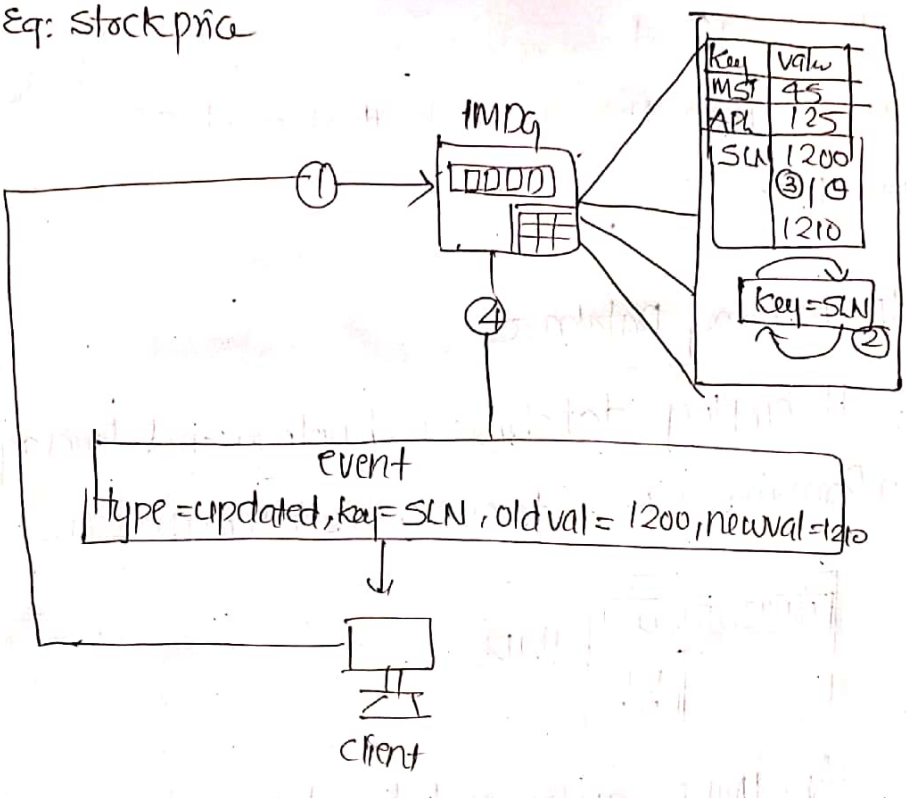


In the figure,

1. An image (a), XML data (b) and customer object (c) are first serialized using a serialization engine
2. They are then stored as key-value pairs in an IMDG.
3. A client requests the customer object via its key.
4. The value is then returned by the IMDG in serialized form.
5. The client then utilizes a serialization engine to deserialize the value to obtain the customer object
6. In order to manipulate the customer object.

IMDGs are heavily used for realtime analytics. This is achieved through continuous querying, also known as active querying.

Eg: stockprice



There are 4 approaches to support read/write performance consistency and simplicity requirements.

1. read-through
2. write-through
3. write-behind
4. refresh-ahead

1. read-through

If a requested value for key is not found in the IMDG, then it is synchronously read from the backend on-disk storage such as database

2. write-through

Any write is written synchronously in a transactional manner to the database. If the write fails, the IMDG's rolled back

3. write-behind

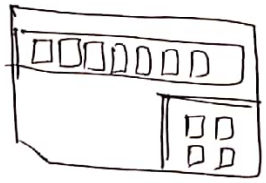
It works in batch-manner. It increases both write & read performance.

4. Refresh-ahead

It is a proactive approach that is it some sets expiry time.

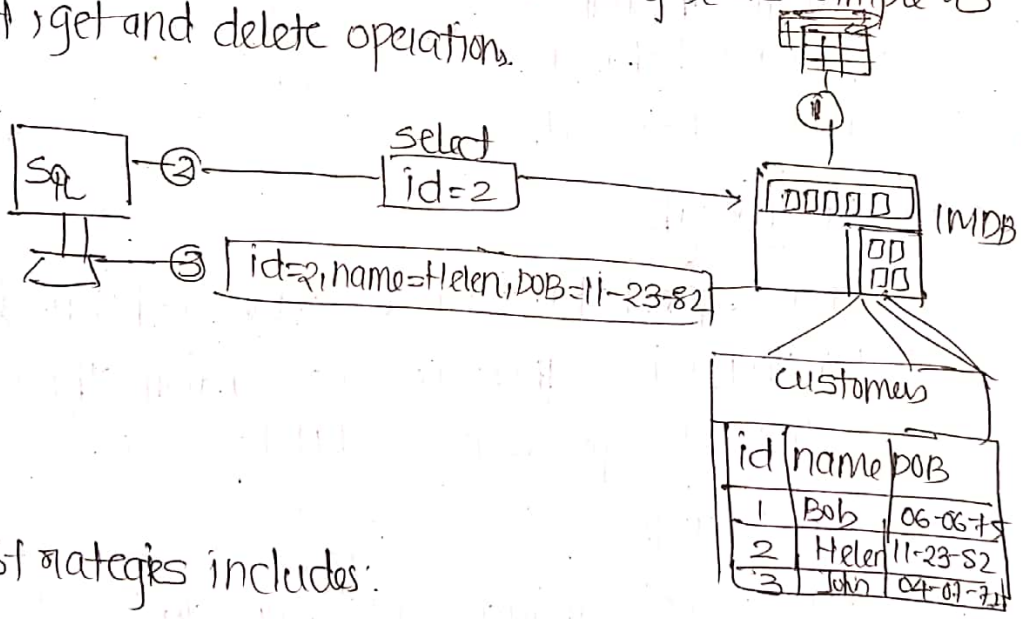
(ii) In-memory Database

It employ database technology and leverage the performance of RAM to overcome latency issues.



An IMDB can be relational in nature for the storage of structured data and unstructured data with familiar SQL language. NoSQL-based IMDBs generally provide API-based access, which may be as simple as put, get and delete operations.

Eg:



Strategies includes:

- Using Non-volatile RAM for storing data permanently
→ Database transaction logs can be periodically stored to a non-volatile medium, such as disk.
→ Snapshot files.
→ An IMDB may leverage sharding & replication

IMDBs are used in real time analytics. It is easy to setup. The main drawback is less scalable the IMDGs.

1. Distributed file system

On-disk storage utilizes low cost hard disk drives for long-term storage. on-disk storage can be implemented via distributed file system or a database. ON-disk storage can be classified in 4:

1. Distributed file system
2. RDBMS database
3. NoSQL database
4. NewSQL database

Distributed file system

It support schema-less data storage. It provides out of box redundancy and high availability by copying data to multiple locations via replication. The distributed file system provides simple, fast access data storage and fast read/write capability.

It is not ideal for datasets comprising a large number of small files as this creates excessive disk seek activity, then slowing down the overall activities. Multiple smaller files are generally combined into a single file to enable optimum storage and processing. This allows the distributed file systems to have increased

performance when data must be accessed in streaming mode with no random reads and writes. It provides an inexpensive storage option for storing large amounts of data over a long period of time. The main drawback is do not provide the ability to search the contents of files