SerDe (Serialization & Deserialization in Hive)

Let's break the hive query & creation.

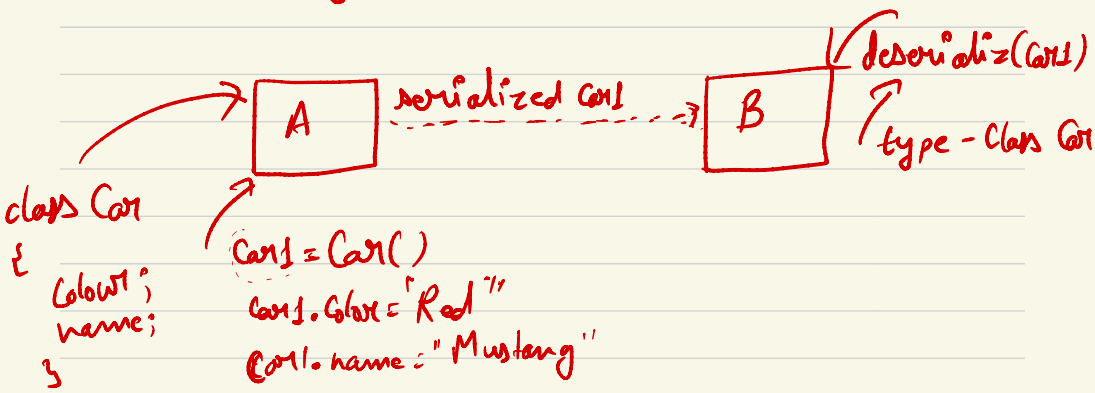
→ Serialization & Deserialization

→ Hive ROW FORMAT

→ Map-Reduce Input/Output

* **Serialization**: It is a process of converting an object into bytes that can be stored in a file or transmitted over the network

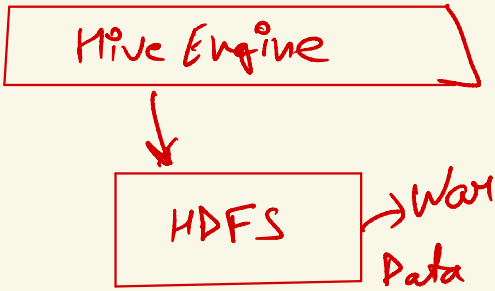
* **Deserialization**: Process of converting bytes into an object.



Hive Query

select {name},{salary} from employee;

Here deserialization of Column/Data will happen



When we run the load command that means, we are loading serialized data in hdfs.

=
> Create table mytable (a string, b string)
> ROW FORMAT delimited
> Stored as textfile;

Row Format: describes the libraries used to convert a given row into column object.

Stored As: describes the input format & output format libraries used by map-reduce to read & write to HDFS.

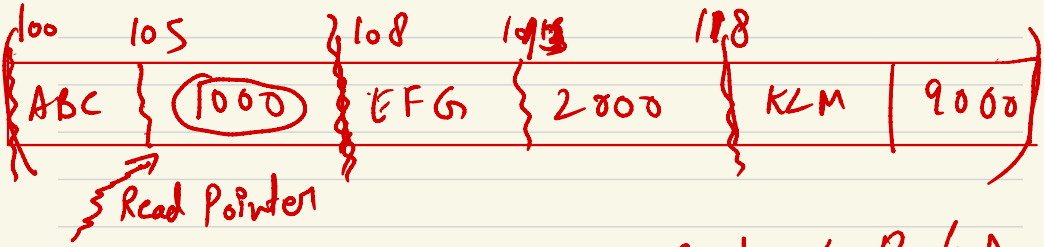
*) Columnar file format / Databases

→ Parquet

→ ORC (Optimized Row Columnar)

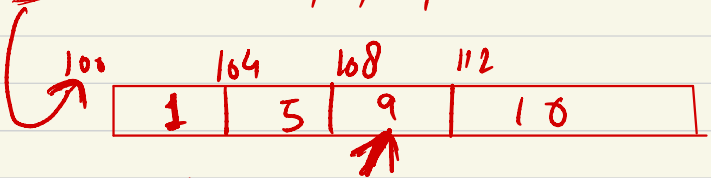
Row Based

employee_table (name, Salary)



int = 4 Bytes

a_list = [1, 5, 9, 10]



alist[2] = a_list + 2*4

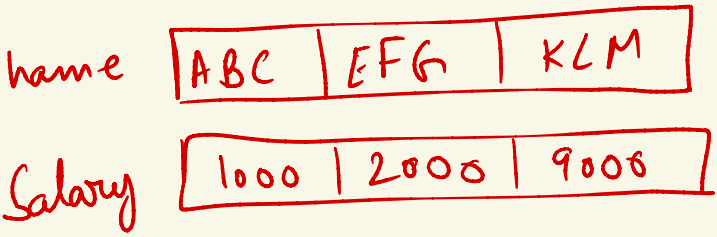
= 100 + 2*4

= 100 + 8 = 108

select sum(Salary) from employee;

In Row Based → Seek Time is super high

Columnar Based



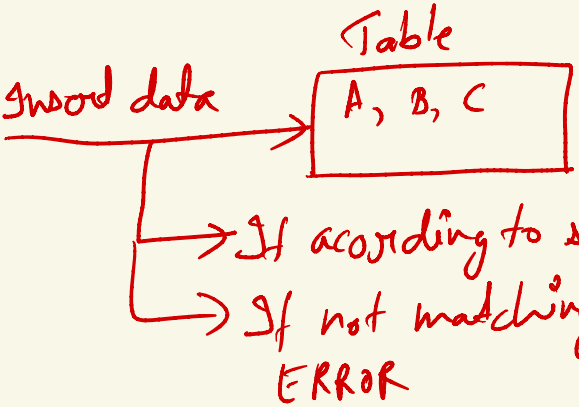
Select sum(Salary) from Employee;

ORC Compression > Parquet Compression

Hive works well with ORC file format.

Spark performs well with parquet file format.

Schema on Write (RDBMS)



*) Validation will happen during the write time or declaration time.

Schema on Read (Hive)