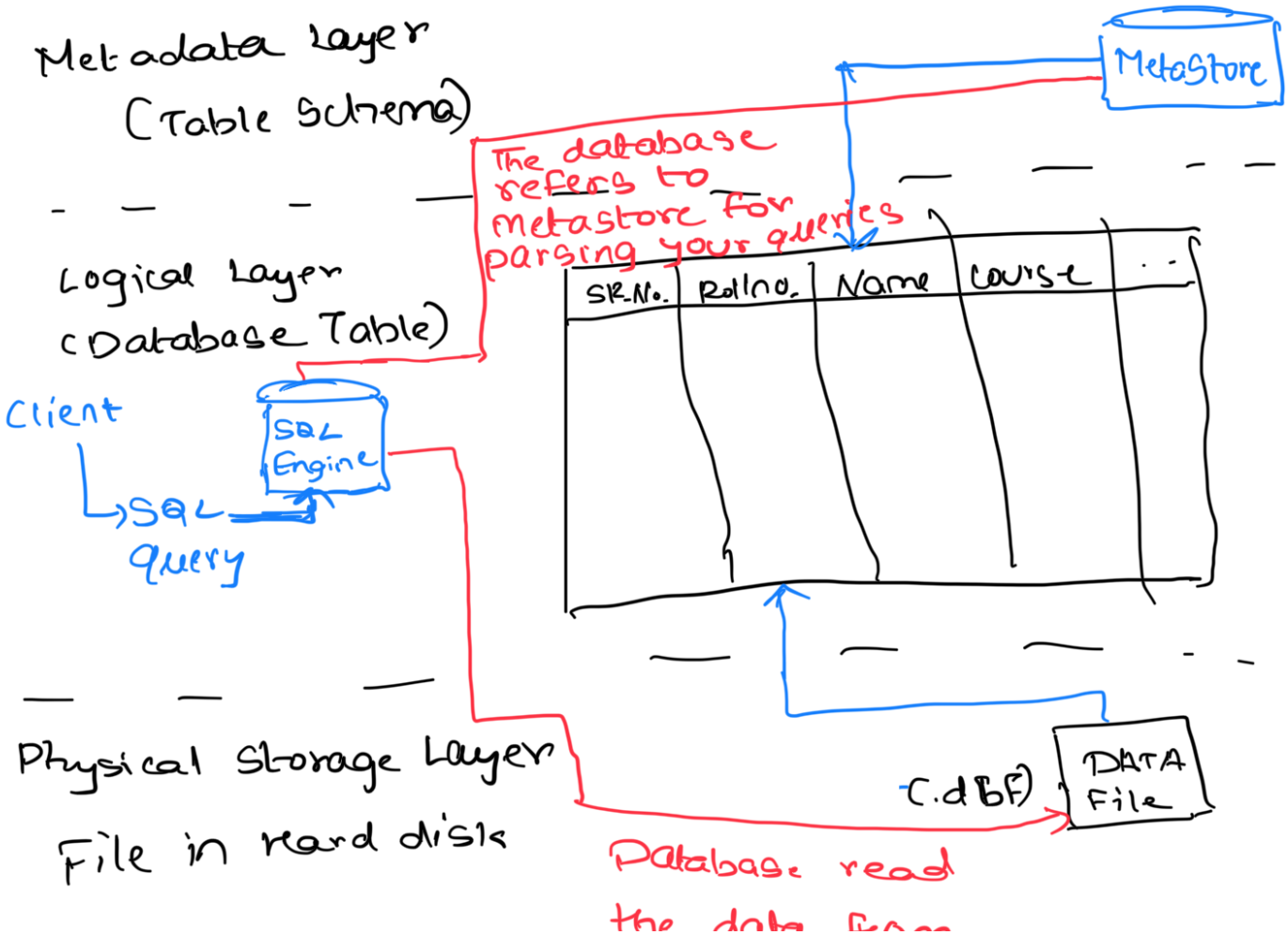
* What is Spark?

→ A Data Processing Platform

⇒ One another Data Processing Platform is Databases. It offers two things in large scale – 1] Table
2] SQL

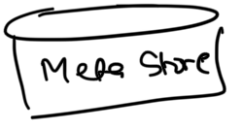
- A Database Table allows you to load the data in the table.
And the data in the table is internally Stored as a .dbf file



.dBF file & process it according to your SQL query & give the result

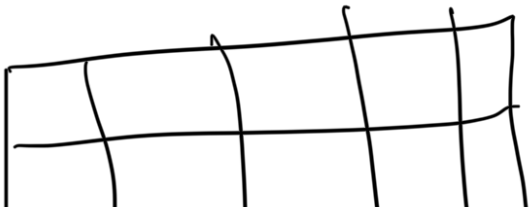
Now let's try Database Analogy with Apache Spark :

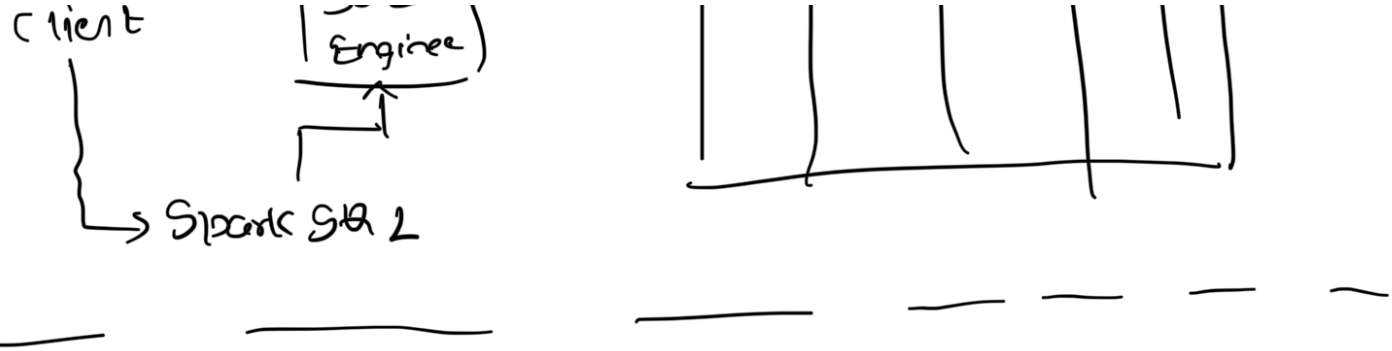
- Apache Spark offers you two ways of data processing :
  1] Spark Database & SQL
  2] Spark Dataframe & Dataframe API
- The first approach is precisely the same as a typical database. You create table & load data into table. Spark table data is internally stored in the data files. But these files are not dbf files. Spark gives you flexibility to choose the file format & supports many file formats



Metadata Layer
Table Schema

Logical Layer
(Table)





Physical Storage Layer

File in Distributed Storage (HDFS, Amazon S3, etc)

CSV, JSON, AVRO, XML, etc. — Data File

⇒ However, Spark does beyond the Tables and SQL to offer Spark Dataframe & Dataframe API.

What is Spark Dataframe?

→ Spark Dataframe is structurally the same as table. However, it does not store any schema information in metadata store. Instead, it has a runtime metadata catalog to store the dataframe schema information.

This catalog is only valid until your application is running. Spark will delete this catalog when your Spark application terminates

Spark Dataframe	Spark Table
- You can create a Spark Dataframe at runtime and keep it in memory until your program terminates.	- Spark tables are permanent. Once created, you will have a table forever.
- Spark Dataframe is a runtime & temporary object, which lives in spark memory and goes away when the application terminates.	- Spark table remains in the system until you drop the table.

The second reason is due to schema-on-read feature:

- In, Spark table we define a schema for the table when creating the table and then we load the data in the table. The data must comply with the table schema, Or you will get an error.
- In Dataframe, we load the data into the Dataframe and tell the schema when [illegible] data. So, dataframe does not

loading the data
have a fixed & predefined schema stored
instead we define the schema, when
we want to read the data from a file
and load it into the Dataframe

So, a Dataframe is always loaded with some data, whereas a Table can be empty.

- And one more thing, Dataframe does not support SQL expressions. You must use Dataframe APIs to process data from a Dataframe

* DataFrame Methods:

1) Actions: Actions are DataFrame operations that kick off a Spark Job execution and return to the Spark driver

2) Transformations: Spark Dataframe transformation produces a newly transformed Dataframe

3) Functions/Methods: Dataframe methods or functions which are not categorized into Actions