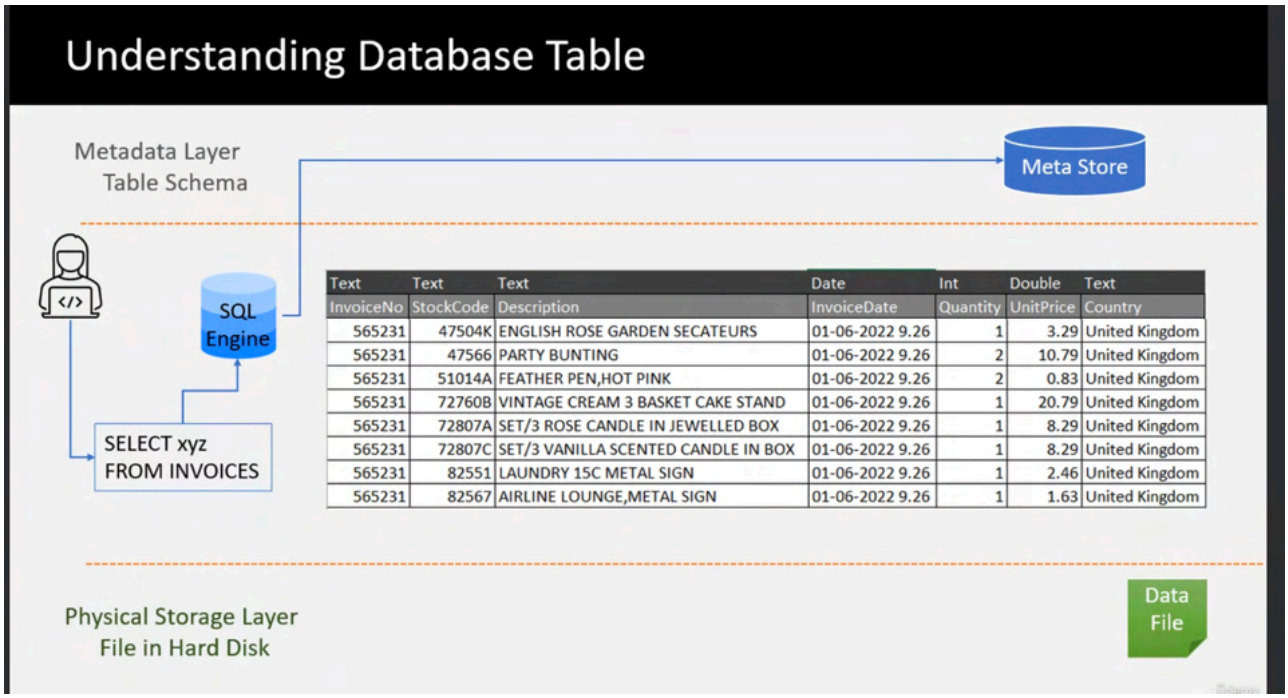
V16 - Introduction to Spark Data Frames

Monday, 20 January 2025 7:40 PM

- **SECTION 3 - V16 - Introduction to Spark Data Frames**
- https://udemy.com/course/apache-spark-programming-in-python-for-beginners/learn/lecture/39877640#overview
  a. ***Spark is a data processing platform***, other data processing platforms are databases.
  b. Databases are most popular and most widely used data processing platforms.
  c. Databases offer two things at high level - ***Tables and SQL language***
  d. A database table allows us to load data in table and data in table is internally stored in a DBF file on disk.
  e. We only care about table and not about DBF files, we see table and we query the table via SQL.
  f. Table contains two things, table schema and table data.
  g. Database table is in logical layer, in physical storage layer if is present as DBF files.
  h. ***Table schema is simply the list of table column names and datatypes***. This schema information is stored in a ***database data dictionary or a meta store***. This is how table is organized.
  i.



Text	Text	Text	Date	Int	Double	Text
InvoiceNo	StockCode	Description	InvoiceDate	Quantity	UnitPrice	Country
565231	47504K	ENGLISH ROSE GARDEN SECATEURS	01-06-2022 9.26	1	3.29	United Kingdom
565231	47566	PARTY BUNTING	01-06-2022 9.26	2	10.79	United Kingdom
565231	51014A	FEATHER PEN,HOT PINK	01-06-2022 9.26	2	0.83	United Kingdom
565231	72760B	VINTAGE CREAM 3 BASKET CAKE STAND	01-06-2022 9.26	1	20.79	United Kingdom
565231	72807A	SET/3 ROSE CANDLE IN JEWELLED BOX	01-06-2022 9.26	1	8.29	United Kingdom
565231	72807C	SET/3 VANILLA SCENTED CANDLE IN BOX	01-06-2022 9.26	1	8.29	United Kingdom
565231	82551	LAUNDRY 15C METAL SIGN	01-06-2022 9.26	1	2.46	United Kingdom
565231	82567	AIRLINE LOUNGE,METAL SIGN	01-06-2022 9.26	1	1.63	United Kingdom

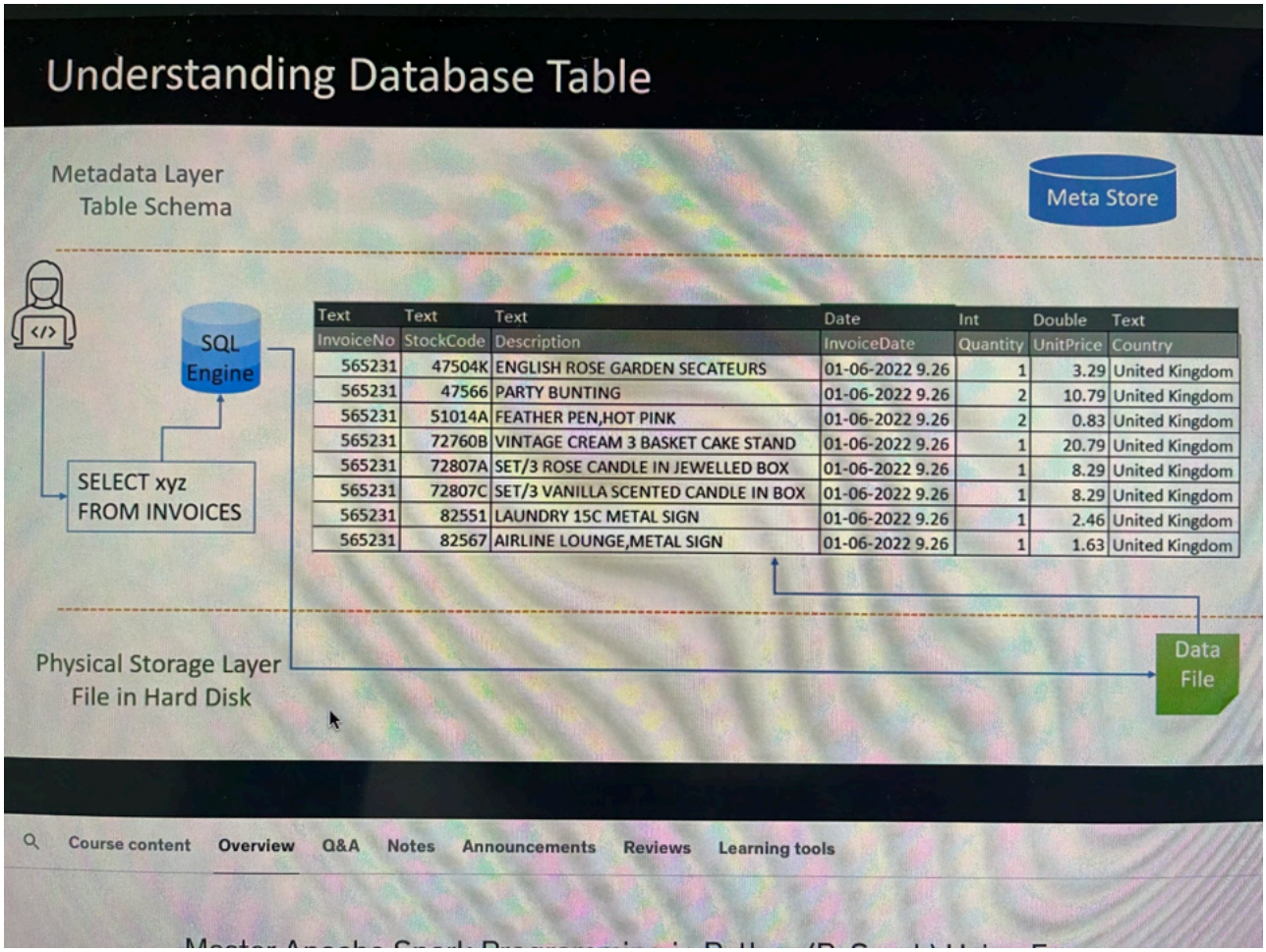
  j. There are in total 3 layers to form a table in a database.
    i. Storage layer stores that table data in a file on disk
    ii. Metadata layer stores the table schema and other imp info

iii. ***Logical layer - presents us with database table and we can execute sql queries on logical table.***

k. ***When we write a sql query and submit it to database sql engine, the DB will refer to metadata store to parse sql query and it will throw a <u>syntax error or analysis error/exception</u> if we are using col name in sql that does not exist in metastore***.

l. If query passes all the ***schema validation*** the database will read data from DBF files, process it according to sql query and display/show you the results on logical layer.

m.



Text	Text	Text	Date	Int	Double	Text
InvoiceNo	StockCode	Description	InvoiceDate	Quantity	UnitPrice	Country
565231	47504K	ENGLISH ROSE GARDEN SECATEURS	01-06-2022 9.26	1	3.29	United Kingdom
565231	47566	PARTY BUNTING	01-06-2022 9.26	2	10.79	United Kingdom
565231	51014A	FEATHER PEN,HOT PINK	01-06-2022 9.26	2	0.83	United Kingdom
565231	72760B	VINTAGE CREAM 3 BASKET CAKE STAND	01-06-2022 9.26	1	20.79	United Kingdom
565231	72807A	SET/3 ROSE CANDLE IN JEWELLED BOX	01-06-2022 9.26	1	8.29	United Kingdom
565231	72807C	SET/3 VANILLA SCENTED CANDLE IN BOX	01-06-2022 9.26	1	8.29	United Kingdom
565231	82551	LAUNDRY 15C METAL SIGN	01-06-2022 9.26	1	2.46	United Kingdom
565231	82567	AIRLINE LOUNGE,METAL SIGN	01-06-2022 9.26	1	1.63	United Kingdom

n. Apache Spark offers two ways of data processing -

i. Spark database and SQL

ii. Spark dataframe and dataframe API

o. First approach is same as typical database system

p. We create table and load data, data is stored instead of DBF files, in different formats like xml, parquet, JSON and ARO, XML. Spark gives flexibility to choose the file format. That is why spark support structured, semi-structured and unstructured data processing.

q. DBF file format and database engine were designed to process only structured data. This was Database limitation.

r. The spark storage layer also supports distributed storage such as HDFS and Cloud storage such as Amazon S3 and Azure ADLS.

So you are not limited to disk storage capacity.
You can use distributed storage and store large data files.

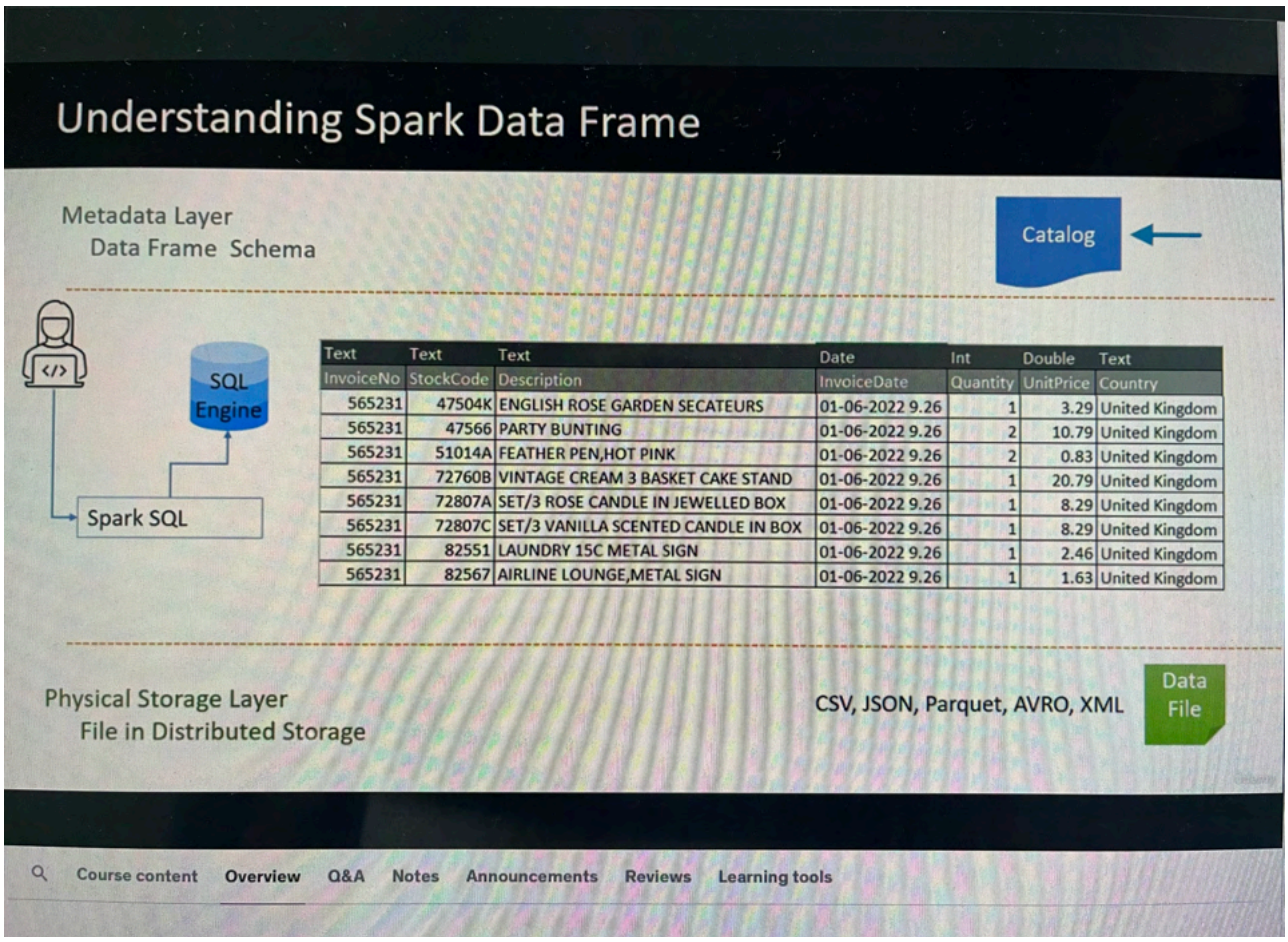
A. ***Spark also has a metadata store for storing table schema information.***
So that part is similar to the databases.
Then Spark also comes with an SQL query engine
and supports standard SQL syntax for processing and querying data from Spark tables.

B. However, Spark goes beyond the Tables and SQL to offer Spark Dataframe and Dataframe API.

C. What is a Spark Dataframe?

D. It is the same as the table without a metadata store.
What does it mean?
Spark Dataframe is structurally the same as the table.
However, it does not store any schema information in the metadata store.

E. Instead, it has a runtime metadata catalog to store the dataframe schema information.
The catalog is similar to the metadata store,
but Spark will create it at the runtime to store schema information in the catalog.
This catalog is only valid until your application is running.
Spark will delete this catalog when your Spark application terminates.



What is the benefit?
Why not store the Dataframe schema in the metadata store?
Why create a runtime metadata catalog?

Two reasons for this.
Spark Dataframe is a runtime object.
and Spark Dataframe supports schema-on-read.

What does it mean?
You can create a Spark Data frame at runtime and keep it in memory until your program terminates.
Once your program terminates, your dataframe is gone.
It is an in-memory object.

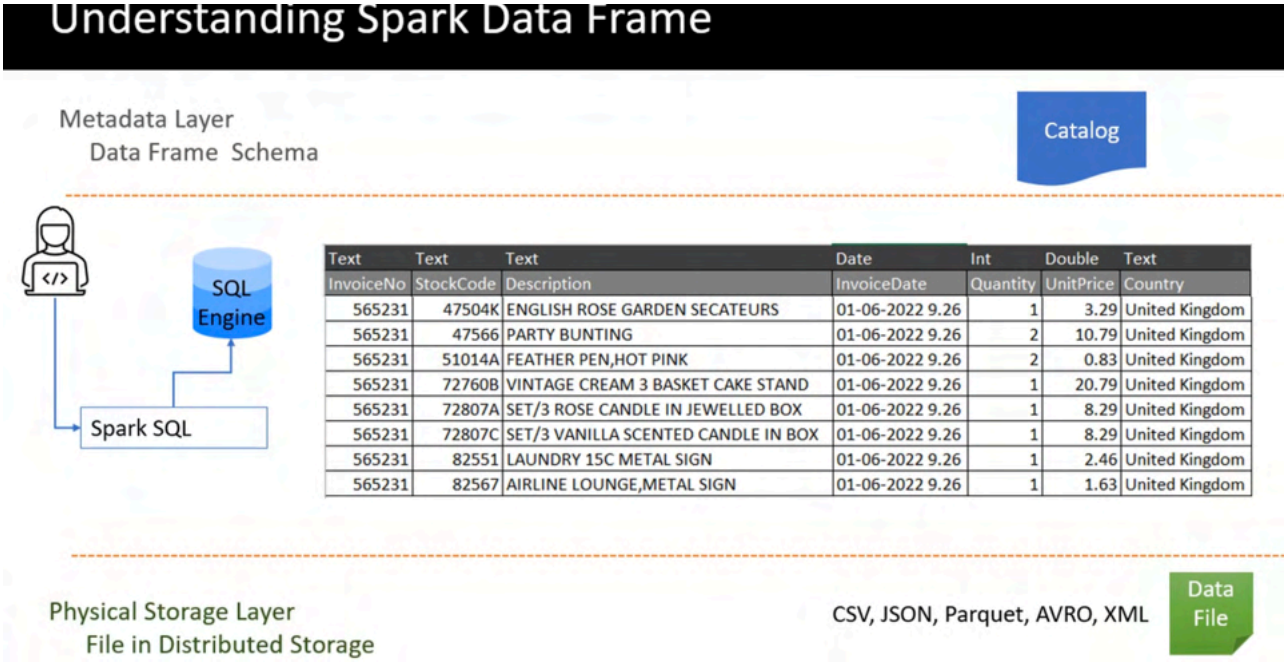
Spark tables are permanent.
Once created, you will have a table forever.
You can drop a table and remove it.

However, ***Spark Dataframe is a runtime and temporary object***
which lives in Spark memory and goes away when the application terminates.
Metadata is also stored in the temporary metadata catalog.

F. ***The second reason is due to the schema-on-read feature.***
Spark Dataframe is designed to support the idea of schema-on-read.
Dataframe does not have a fixed and predefined schema stored in the metadata store.
Instead, we define the schema when we want to read the data from a file and load it into the Dataframe.
So the difference is straightforward.
We define a schema for the table when creating a table.
Then we load data into the table.
The data must comply with the table schema, or you will get an error.

However, Dataframe is different.
We load the data into a Dataframe and tell the schema when loading the data.
And Spark will read the file, apply the schema at the time of reading,
create the Dataframe using the schema and load the data.
So a Dataframe is always loaded with some data, whereas a Table can be empty.

G.



Text	Text	Text	Date	Int	Double	Text
InvoiceNo	StockCode	Description	InvoiceDate	Quantity	UnitPrice	Country
565231	47504K	ENGLISH ROSE GARDEN SECATEURS	01-06-2022 9.26	1	3.29	United Kingdom
565231	47566	PARTY BUNTING	01-06-2022 9.26	2	10.79	United Kingdom
565231	51014A	FEATHER PEN,HOT PINK	01-06-2022 9.26	2	0.83	United Kingdom
565231	72760B	VINTAGE CREAM 3 BASKET CAKE STAND	01-06-2022 9.26	1	20.79	United Kingdom
565231	72807A	SET/3 ROSE CANDLE IN JEWELLED BOX	01-06-2022 9.26	1	8.29	United Kingdom
565231	72807C	SET/3 VANILLA SCENTED CANDLE IN BOX	01-06-2022 9.26	1	8.29	United Kingdom
565231	82551	LAUNDRY 15C METAL SIGN	01-06-2022 9.26	1	2.46	United Kingdom
565231	82567	AIRLINE LOUNGE,METAL SIGN	01-06-2022 9.26	1	1.63	United Kingdom

H. You can use SQL on the table.

Dataframe does not support SQL expressions.

We must use Dataframe APIs to process data from a Dataframe.

Since the table and dataframe are structurally the same, you can convert them to each other.

You can create a table and use SQL or convert a table into a Dataframe and use Dataframe API on the same table.

I.