https://community.cloud.databricks.com/?o=6951507992936870#

Clusters -> Create Cluster

Workspace -> Right Click -> Create -> Attach Cluster

-> Import Data

Parquet: What is Apache Parquet?

Pivot Table: SQL Pivot: Converting Rows to Columns - The Databricks Blog

Widgets: Widgets — Databricks Documentation

CREATE TABLE

```
DROP TABLE IF EXISTS outdoorProducts;
CREATE TABLE outdoorProducts (
            invoiceNo       STRING,
             stockCode     STRING,
             description STRING,
             quantity INT,
             invoiceDate STRING,
             unitPrice       DOUBLE,
             customerID INT,
             countryName STRING
) USING csv OPTIONS (
  PATH "/mnt/training/online_retail/data-001/data.csv",
  header "true"
);
```

CREATE TEMPORARY VIEW

```
CREATE
OR REPLACE TEMPORARY VIEW sales AS
SELECT
            stockCode,
            quantity,
            unitPrice,
            quantity*unitPrice AS totalAmount,
            countryName
FROM
            outdoorProducts
WHERE
            quantity > 0;
```

DESCRIBE tabela

DESCRIBE EXTENDED tabela

CACHE TABLE tabela

UNCACHE TABLE IF EXISTS tabela

PARTITIONS

```
DROP TABLE IF EXISTS bikeShare_partitioned;
CREATE TABLE bikeShare_partitioned
PARTITIONED BY (p_hr)
  AS
SELECT
  instant,
  workingday,
  weathersit,
  temp
FROM
  bikeShare
```

SHOW PARTITIONS p_hr

EXPLODE NESTED OBJECT (LISTS)

```
SELECT EXPLODE (source)
FROM DCDataRaw;
```

(CTE) COMMON TABLE EXPRESSIONS - TEMPORARY

```
WITH ExplodeSource  -- specify the name of the result set we will query
AS
(                -- wrap a SELECT statement in parentheses
  SELECT           -- this is the temporary result set you will query
             dc_id,
```

```
            to_date(date) AS date,
            EXPLODE (source)
  FROM
            DCDataRaw
)
SELECT          -- write a select statment to query the result set
          key,
          dc_id,
          date,
          value.description,
          value.ip,
          value.temps,
          value.co2_level
FROM            -- this query is coming from the CTE we named
          ExplodeSource;
```

(CTAS) CREATE TABLE AS SELECT - PERMANENT

```
DROP TABLE IF EXISTS DeviceData;
CREATE TABLE DeviceData
USING parquet
WITH ExplodeSource                  -- The start of the CTE from the last cell
AS
  (
  SELECT
          dc_id,
          to_date(date) AS date,
          EXPLODE (source)
  FROM
         DCDataRaw
  )
SELECT
          dc_id,
          key device_type,
          value.temps,
          value.co2_level

FROM ExplodeSource;

SELECT * FROM DeviceData;
```

SAMPLE TABLE

SELECT * FROM outdoorProductsRaw **TABLESAMPLE (5 ROWS)**
SELECT * FROM outdoorProductsRaw **TABLESAMPLE (2 PERCENT)** ORDER BY InvoiceDate

---

CHECK NULL VALUES

SELECT count(*) FROM outdoorProductsRaw WHERE Description **IS NULL**

REPLACE NULL VALUES

SELECT
    COALESCE(Description, "Misc"**)** AS Description,
FROM
    outdoorProductsRaw

SPLIT STRING AND RETURN ARRAY

SELECT
    SPLIT(InvoiceDate, **"/")[0]** month,
    SPLIT(InvoiceDate, **"/")[1]** day,
    SPLIT(SPLIT(InvoiceDate, **" ")[0], "/")[2]** year
FROM
    outdoorProductsRaw

INSERT CHARACTER ON THE LEFT OF A STRING

SELECT
    LPAD(month, 2, 0) AS month,
    LPAD(day, 2, 0) AS day
FROM outdoorProducts

CONCATENATING

```
SELECT
        CONCAT_WS("/", month, day, year)
FROM outdoorProducts
```

CHANGE TO DATA TYPE

```
SELECT
        to_date(sDate, "MM/dd/yy") date,
        CAST(UnitPrice AS DOUBLE)
FROM
        standardDate
```
