Recap:

Joins

1. Inner
2. Left, leftouter
3. Right , rightouter
4. Full, fullouter
5. **Leftsemi**
6. **Leftanti**
7. Cross

========================

Write the empdf and deptdf dataframes into retail database and perform join operations

=======================================

Join optimization 👍

1. Broadcast join
2. Sort merge join
3. Shuffle Hash join

Student performance :

Fees

Attendance

Univ

College

Stream

Semester

Faculty

Subjects

**Marks**

**Cgpa --measure , metric, fact, KPI -key performance Indicator**

Extra

==================================

Beverages

Soda

Pepsi

Diet pepsi

Asia pacific

India

SI

Tel

Hyderabad

GB

ABC store

There is an upper limit in terms of records as well. We can't broadcast more than 512m records. So its either 512m records or 8GB which ever limit hits first

spark.sql.autoBroadcastJoinThreshold

By default => 10mb

=======================================

MEMORY management :

Empdf

Salesdf

salesdf.cache()

salesdf.show()

empdf.show()

-----cut the files from the source ----

salesdf.show()

empdf.show()

Project task 1:

Problem definition /Statement :

KYD --Know Your Data

No of columns , description , primary key , foreign key

No of rows , missing , null , mismatch , delimiter

Cleansing and Scrubbing

Airlines :

Data source : csv (airport) === spark --file

Hdfs (flights) → Hive External ⇒ spark

RDB (airlines) ⇒ Mysql ==>spark

Loan : (using pandas split the data into three dataframe and write it to separate file

Csv ⇒ spark ⇒ df⇒ filter ---> 2001

Hdfs → hive external table ⇒spark ⇒ 2002

Mysql → JDBC ==spark ⇒2003

Retail :

2 files each ,load it to the following sources

Data source : csv === spark --file

Hdfs → Hive External ⇒ spark

RDB ⇒ Mysql ==>spark

Covid :

Hdfs (worldometer) → Hive External ⇒ spark

RDB (Countrywise) ⇒ Mysql ==>spark

Altenate fuel :

**Project Task : Day 10**

Problem definition /Statement :

KYD --Know Your Data

No of columns , description , primary key , foreign key

No of rows , missing , null , mismatch , delimiter

Cleansing and Scrubbing

Airlines :

Data source : csv (airport) === spark --file

Hdfs (flights) → Hive External ⇒ spark

RDB (airlines) ⇒ Mysql ==>spark

Loan : (using pandas split the data into three dataframe and write it to separate file

1947--2000 Csv ⇒ spark ⇒ df⇒ filter ---> 2001

2001-2020 Hdfs → hive external table ⇒spark ⇒ 2002

Latest Mysql → JDBC ==spark ⇒2003

Retail :

2 files each ,load it to the following sources

Data source : csv === spark --file

Hdfs → Hive External ⇒ spark

RDB ⇒ Mysql ==>spark

Covid :

https://github.com/CSSEGISandData/COVID-19/tree/master/csse\_covid\_19\_data/csse\_covid\_19\_daily\_reports

Data source : csv (2020) === spark --file

Hdfs (2021) → Hive External ⇒ spark

RDB (2022) ⇒ Mysql ==>spark

Altenate fuel :

https://afdc.energy.gov/data\_download

Data source : Alternate fuelstations : csv === spark --file

Light duty vehicles : Hdfs → Hive External ⇒ spark

Medium and Heavy => RDB ⇒ Mysql ==>spark