Homework 3

Data Science And Database Technology

The following relations are given (primary keys are underlined):

ACCOMODATION(CodA, NumberOfGuests, Address, City, Region)
SERVICE(CodS, ServiceName, ServiceType)
ACCOMODATION-HAS-SERVICE(CodA, CodS)
USER(CodU, FirstName, Surname, BusinessAccount, BirthDate, Address, City, Region)
BOOKING(CodA, StartDate, CodU, EndDate)

Assume the following cardinalities:

- card(ACCOMODATION) $=10^5$ tuples,
  distinct values of Region = 20
- card(SERVICES)$=10^2$ tuples,
  distinct values of ServiceType = 20
- card(ACCOMODATION-HAS-SERVICE)$=10^6$ tuples,
- card(USER)$=10^4$ tuples,
  MIN(DATE(BirthDate)) = 1/1/1930,
  MAX(DATE(BirthDate)) = 31/12/2009,
  distinct values of Region = 20,
  distinct values of BusinessAccount = 2 ("True", "False")
- card(BOOKING)$=10^7$ tuples,
  MIN(Date) = 1/9/2017, MAX(Date)) = 31/08/2020

Furthermore, assume the following reduction factor for the group by condition:

- Having COUNT(Distinct StartDate)>1 $\approx$ 1/10

Consider the following SQL query:

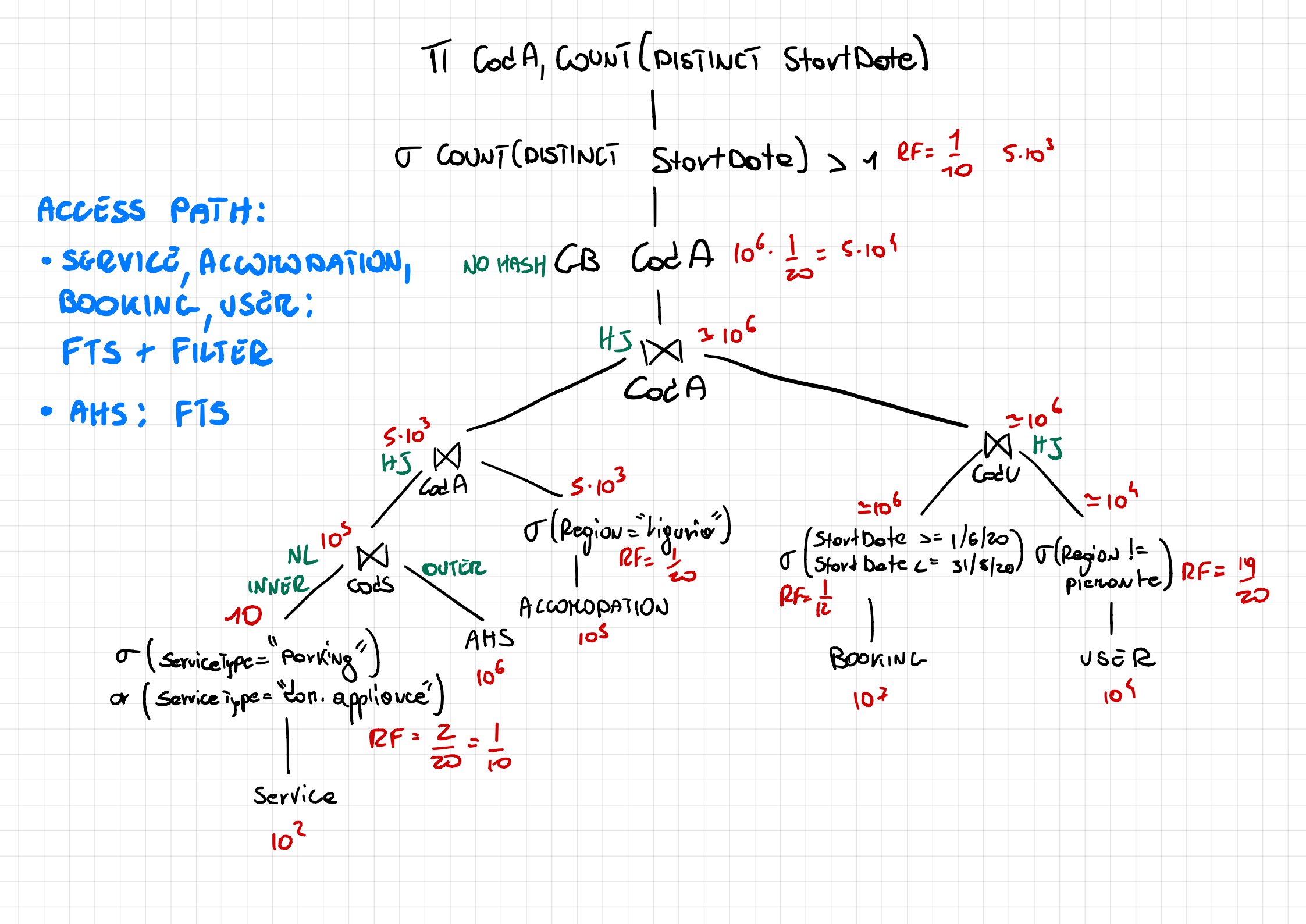
```
select A.CodA, count(Distinct StartDate)
from SERVICE S, ACCOMODATION-HAS-SERVICE AHS,
ACCOMODATION A, BOOKING B, USER U
where S.CodS=AHS.CodS and A.CodA=AHS.CodA and
U.CodU=B.CodU and B.CodA=A.CodA
and (S.ServiceType=”Parking” or S.ServiceType=”domestic
appliances”)
and  A.Region=”Liguria” and B.StartDate>=1/5/20 and
B.StartDate<=31/8/20
and U.Region<>”Piemonte”
group by CodA
Having COUNT(Distinct StartDate)>1
```

Homework tasks

For the SQL query:

1. Report the corresponding algebraic expression and specify the cardinality of each node (representing an intermediate result or a leaf). If necessary, assume a data distribution. Also analyze the GROUP BY anticipation.

2. Select one or more secondary physical structures to increase query performance. Justify your choice and report the corresponding execution plan (join orders, access methods, etc.).


ACCESS PATH:
• SERVICE, ACCOMODATION, BOOKING, USER: FTS + FILTER
• AHS: FTS
Π CodA, COUNT(DISTINCT StartDate)
σ COUNT(DISTINCT StartDate) > 1
RF = 1/10
5·10^3
NO HASH GB CodA
10^6 · 1/20 = 5·10^4
HJ
⋈
CodA
≈10^6
5·10^3
HJ
⋈
CodA
5·10^3
≈10^6
HJ
⋈
CodU
≈10^4
NL
INNER
10^5
⋈
CodS
OUTER
σ (Region = "liguria")
RF = 1/20
ACCOMODATION
10^5
≈10^6
σ (StartDate >= 1/6/20
Start Date <= 31/8/20)
RF = 1/12
σ (Region != piemonte)
RF = 19/20
σ (ServiceType = "Parking"
or (ServiceType = "Con. appliance")
RF = 2/20 = 1/10
AHS
10^6
BOOKING
10^7
USER
10^4
Service
10^2

INDICES:

- SERVICE: NO
- AHS: NO
- ACCOMODATION: SECONDARY HASH INDEX ON "Region"
- BOOKING: SECONDARY B+ TREE INDEX ON "StartDate"
- USER: NO

ACCESS PATH (INDEX):

- ACCOMODATION: INDEX FULL SCAN
- BOOKING: INDEX RANGE SCAN

WITH GB ANTICIPATION