Question
Nicky Zhu · Feb 12, 2023

How to optimize slow count distinct SQL

Hi guys,

I'm working on a project with my client.
They have a visit table which has about 7,000,000 records. The table is used in a random search page witch holds 20+ conditions to be combined.
The table is defined as below:

```
CREATE TABLE SQLUser.ST_SEARCH_VISIT (

HQ_ORG_CODE VARCHAR(32),

HQ_ORG_NAME VARCHAR(32),

TENANT_ID VARCHAR(32),

GROUP_PATIENT_SN VARCHAR(32),

ORG_PATIENT_SN VARCHAR(32),

NAME VARCHAR(64),

SEX_CODE VARCHAR(2),

SEX_NAME VARCHAR(16),

BIRTH_DATE DATE,

MARRY_CODE VARCHAR(2),

MARRY_NAME VARCHAR(16),

IDENTIFY_TYPE_CODE VARCHAR(5),

IDENTIFY_TYPE_NAME VARCHAR(32),

IDENTIFY_NUMBER VARCHAR(50),

BLOOD_TYPE_CODE VARCHAR(2),

BLOOD_TYPE_NAME VARCHAR(12),

MOBILE VARCHAR(50),

MAILING_ADDRESS VARCHAR(127),

VISIT_SERIAL_NO VARCHAR(64),

TABLE_FLAG VARCHAR(1),
```

```
VISIT_TYPE_CODE VARCHAR(24),

VISIT_TYPE_NAME VARCHAR(64),

VISIT_DEPT_CODE VARCHAR(32),

VISIT_DEPT_NAME VARCHAR(64),

IN_OUT_VISIT_NO VARCHAR(32),

VISIT_TIME TIMESTAMP,

DISCHARGE_TIME TIMESTAMP,

IN_HOSPITAL_TIME TIMESTAMP,

ICD_CODE VARCHAR(20),

ICD_NAME VARCHAR(64),

HEALTH_CARD_NO VARCHAR(32),

HEALTH_CARD_TYPE VARCHAR(64),

SURG_CODE VARCHAR(20),

SURG_NAME VARCHAR(64),

AGE VARCHAR(10),

DISEASE_CODE INTEGER,

DISEASE_NAME VARCHAR(64),

LastChangedTime TIMESTAMP,

LastCreateTime TIMESTAMP,

GROUPORG VARCHAR(50)

);

CREATE INDEX
 BITMAP_INDEX_ST_SEARCH_VISIT_BLOOD_TYP
E_CODE ON SQLUser.ST_SEARCH_VISIT (BLOOD_TYPE_CODE);

CREATE INDEX
 BITMAP_INDEX_ST_SEARCH_VISIT_HQ_ORG_CODE ON SQLUser.ST_SEARCH_VISIT (HQ_ORG_CODE);

CREATE INDEX
 BITMAP_INDEX_ST_SEARCH_VISIT_IDENTI
FY_TYPE_CODE ON SQLUser.ST_SEARCH_VISIT (IDENTIFY_TYPE_CODE);

CREATE INDEX
 BITMAP_INDEX_ST_SEARCH_VISIT_MARRY_CODE ON SQLUser.ST_SEARCH_VISIT (MARRY_CODE);

CREATE INDEX
 BITMAP_INDEX_ST_SEARCH_VISIT_SEX_CODE ON SQLUser.ST_SEARCH_VISIT (SEX_CODE);
```

```
CREATE INDEX
 BITMAP_INDEX_ST_SEARCH_VISIT_TABLE_FLAG ON SQLUser.ST_SEARCH_VISIT (TABLE_FLAG);

CREATE INDEX
 BITMAP_INDEX_ST_SEARCH_VISIT_VISIT_DEP
T_CODE ON SQLUser.ST_SEARCH_VISIT (VISIT_DEPT_CODE);

CREATE INDEX
 BITMAP_INDEX_ST_SEARCH_VISIT_VISIT_TYP
E_CODE ON SQLUser.ST_SEARCH_VISIT (VISIT_TYPE_CODE);

CREATE INDEX INDEX_PI_PATIENT_INFO_HQ_ORG_CODE ON
 SQLUser.ST_SEARCH_VISIT (HQ_ORG_CODE,GROUP_PATIENT_SN);

CREATE INDEX INDEX_ST_SEARCH_VISIT_BIRTH_DATE ON
 SQLUser.ST_SEARCH_VISIT (BIRTH_DATE);

CREATE INDEX
 INDEX_ST_SEARCH_VISIT_DISCHARGE_TIME ON SQLUser.ST_SEARCH_VISIT (DISCHARGE_TIME);

CREATE INDEX
 INDEX_ST_SEARCH_VISIT_GROUP_PATIENT_SN ON
 SQLUser.ST_SEARCH_VISIT (GROUP_PATIENT_SN);

CREATE INDEX
 INDEX_ST_SEARCH_VISIT_IDENTIFY_NUMBER ON SQLUser.ST_SEARCH_VISIT (IDENTIFY_NUMBER);

CREATE INDEX
 INDEX_ST_SEARCH_VISIT_IN_HOSPITAL_TIME ON
 SQLUser.ST_SEARCH_VISIT (IN_HOSPITAL_TIME);

CREATE INDEX
 INDEX_ST_SEARCH_VISIT_IN_OUT_VISIT_NO ON SQLUser.ST_SEARCH_VISIT (IN_OUT_VISIT_NO);

CREATE INDEX INDEX_ST_SEARCH_VISIT_MOBILE ON SQLUser.ST_SEARCH_VISIT (MOBILE);

CREATE INDEX INDEX_ST_SEARCH_VISIT_NAME ON SQLUser.ST_SEARCH_VISIT (NAME);

CREATE INDEX
 INDEX_ST_SEARCH_VISIT_ORG_PATIENT_SN ON SQLUser.ST_SEARCH_VISIT (ORG_PATIENT_SN);

CREATE INDEX
 INDEX_ST_SEARCH_VISIT_VISIT_SERIAL_NO ON SQLUser.ST_SEARCH_VISIT (VISIT_SERIAL_NO);

CREATE INDEX INDEX_ST_SEARCH_VISIT_VISIT_TIME ON
 SQLUser.ST_SEARCH_VISIT (VISIT_TIME);

CREATE INDEX IdxGpsnorg ON SQLUser.ST_SEARCH_VISIT (GROUP_PATIENT_SN,HQ_ORG_CODE);

CREATE INDEX IdxMapOrg ON SQLUser.ST_SEARCH_VISIT (HQ_ORG_CODE);

CREATE INDEX IdxVisitorg ON SQLUser.ST_SEARCH_VISIT (VISIT_SERIAL_NO,HQ_ORG_CODE);

CREATE INDEX LastChangedTimeIndex ON SQLUser.ST_SEARCH_VISIT (LastChangedTime);

CREATE INDEX LastCreateTimeIndex ON SQLUser.ST_SEARCH_VISIT (LastCreateTime);

CREATE INDEX idxGO ON SQLUser.ST_SEARCH_VISIT (GROUPORG);
```

Now we have a slow SQL as below:

```
select count(1) from

(select distinct by
 (b.VISIT_SERIAL_
NO,b.HQ_ORG_CODE) b.VISIT_SERIAL_NO,b.HQ_ORG_CODE  from SQLUser.ST_SEARCH_VISIT b )
```

Which is used to count distinct visits in about 20 hospitals.

This SQL will take more than 10 seconds to finish on their server.

I did a test on my virtual machines with the same table but with populated data(6,000,000 records) then it will took about 16 seconds to finish the search.

I tried to optimize the sql with the followings

------about 12.5s-------

```
select count(1) from (
    select distinct by (b.VISIT_SERIAL_NO,b.HQ_ORG_CODE) b.VISIT_SERIAL_NO,b.HQ_ORG_C
ODE  from %PARALLEL ST_SEARCH_VISIT b
    group by b.HQ_ORG_CODE
)
```

------about 24s-------

```
select SUM(OrgSum) from (

select HQ_ORG_CODE,count(1) as OrgSum from %PARALLEL (

select distinct by (VISIT_SERIAL_NO,HQ_ORG_CODE) VISIT_SERIAL_NO,HQ_ORG_CODE

From %PARALLEL ST_SEARCH_VISIT

Group by HQ_ORG_CODE

)

group by HQ_ORG_CODE

)
```

So what optimizations could be done to make it faster? thanks.

#Performance #SQL #InterSystems IRIS for Health