2024-04-03 PostgreSQL Localization

현행 문서: PostgreSQL Localization

1 개요

로케일(Locale) 지원이란 영문자, 정렬, 숫자, 형식 등 문화적인 기본 설정과 관련된 지원을 말한다. 로케일 지원은 initdb를 사용하여 데이터베이스 클러스터가 생성되면 자동으로 초기화된다. initdb는 기본적으로 실행 환경의 로케일 설정을 사용하여 데이터베이스 클러스터를 초기화하므로, 데이터베이스 클러스터에서 시스템이 이미 원하는 로케일로 설정된 경우 사용자가 특별히 할 일은 없다.

일부 로케일 카테고리는 데이터베이스가 생성될 때 고정된 값이어야 한다. 서로 다른 데이터베이스에 대해 각각 다른 설정을 할 수 있지만 데이터베이스가 생성된 다음에는 해당 데이터베이스에 대한 설정을 변경할 수 없다. 이러한 카테고리 중 하나가 LC_COLLATE다.

2 동작

로케일 설정은 다음과 같은 SQL 기능에 영향을 준다.

- Order by를 사용한 쿼리에서 정렬 순서 또는 텍스트 데이터에서 표준 비교 연산자
- upper 및 lower, initcap 함수
- 패턴 일치 연산자(Like, Similar to 및 POSIX 스타일 정규식). 대소문자 비구분 일치 및 문자 클래스 정규식에 의한 문자 분류에 모두 영향을 미치는 로케일
- to_char 계열 함수
- Like 절을 사용한 인덱스 사용 능력

PostgreSQL에서 C 또는 POSIX가 아닌 다른 로케일을 사용할 때 단점은 성능이다. 문자 처리가 느려지고 Like에서 사용되는 일반 인덱스가 방지된다. 이러한 이유로 실제로 필요한 경우에만 로케일을 사용해야 한다.

3 테스트

LC_COLLATE의 값에 따라 어떠한 차이점이 있는지 한글 정렬과 인덱스 사용 능력을 비교해보았다.

database의 lc collate를 아래와 같이 만들었다.

	datname	datcollate
1	enusdb	en_US.UTF-8
2	kokrdb	ko_KR.UTF-8
3	cdb	C

한글 정렬 테스트를 위해 한글로 입력된 공항 이름 컬럼을 사용했고, Like절 인덱스 테스트를 위해 임직원 이름 컬럼에 인덱스를 생성했다. Like절 인덱스는 검색 입력값으로 시작하는 Like절, 즉 ‘입력값%’일 경우 사용 가능하고 ‘%입력값%’ 또는 ‘%입력값’의 형식은 사용 불가능하다.

3-1 LC_COLLATE = ‘en_US.UTF-8’

1. Order By로 한글 정렬

	korean_airport_name
1	괌, 괌
2	렌, 프랑스
3	포, 프랑스
4	홍콩
5	다윈 , 호주
6	리즈 브래드포드, 영국
7	고치, 인도
8	난디, 피지
9	난징, 중국
10	다롄, 중국
11	더럼, 영국
12	델리, 인도
13	리마, 페루
14	몰타, 몰타
15	뮌헨, 독일
16	방콕, 태국

한글 정렬이 정상적으로 수행되지 않음을 확인할 수 있다.

2. 인덱스가 생성된 컬럼에 Like절 조건을 걸고 조회 쿼리 실행계획 확인

Output Plan

Operation	Params	Rows	Actual Rows	Total Cost	Actual Total Time
Select					
Full Scan (Seq Scan)	table: tpsnl;	4	45	2582.07	11.729

풀스캔으로 수행하는 것으로 확인 된다.

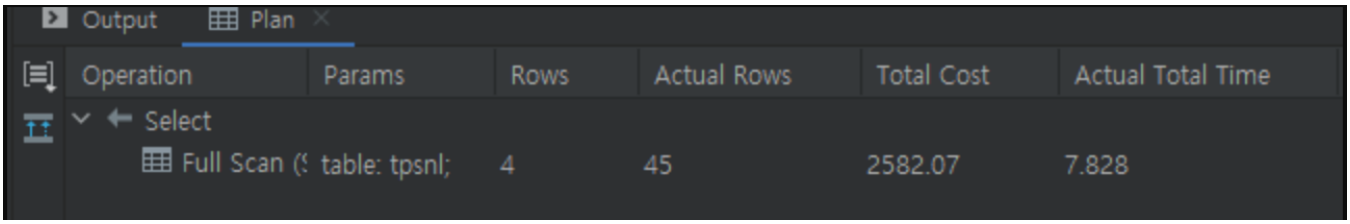
3-2 LC_COLLATE = ‘ko_KR.UTF-8’

1. Order By로 한글 정렬 수행

	korean_airport_name
1	가고시마, 일본
2	가오슝
3	걸프포트/빌록시, MS, 미국
4	게인스빌, FL, 미국
5	고마쓰, 일본
6	고센버그, 스웨덴
7	고치, 인도
8	골드코스트, 호주
9	과달라하라, 멕시코
10	과야킬, 에콰도르
11	괌, 괌
12	광저우, 중국
13	광주, 대한민국
14	그라나다, 스페인
15	그라즈, 오스트리아
16	그란카나리아, 스페인

한글 정렬이 정상적으로 수행됨을 확인할 수 있다.

2. 인덱스가 생성된 컬럼에 Like절 조건을 걸고 조회 쿼리 실행계획 확인



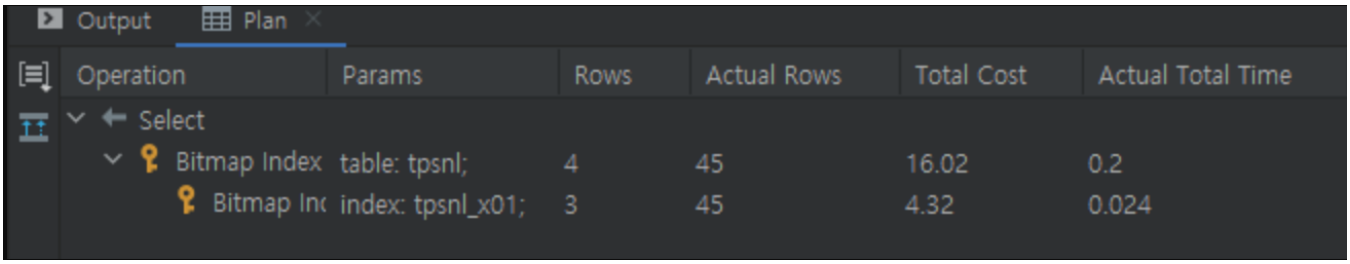
풀스캔으로 수행하는 것으로 확인 된다.

3-3 LC_COLLATE = 'C'

1. Order By로 한글 정렬 수행

한글 정렬이 정상적으로 수행됨을 확인할 수 있다.

2. 인덱스가 생성된 컬럼에 Like절 조건을 걸고 조회 쿼리 실행계획 확인



인덱스를 이용하여 수행하는 것으로 확인 된다.

4 결론

LC COLLATE값에 따른 테스트 결과는 다음과 같다.

	en_US.UTF-8	ko_KR.UTF-8	C
한글 정렬	비정상 수행	정상 수행	정상 수행
Like절 인덱스 사용	불가능	불가능	가능

LC_COLLATE의 값이 en_US.TUF-8인 경우 한글 정렬과 Like절 인덱스 사용에 문제가 있었다. ko_KR.UTF을 사용하면 한글 정렬에는 문제가 없지만 Like 절 인덱스 사용이 불가능하다.

따라서 특정 로케일을 지정해서 사용해야 될 경우가 아니라면 LC_COLLATE = 'C'로 설정하는 것이 좋다.

5 참고 문헌

https://www.postgresql.org/docs/current/charset.html