



데이터 플랫폼 이론-03

엄진영

• 빅쿼리, 빅데이터 저장 및 분석 플랫폼

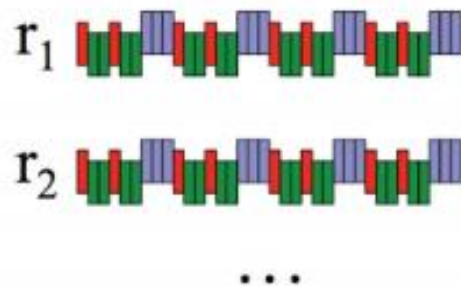
- 빅쿼리는 페타바이트급 데이터 저장 및 분석용 클라우드 서비스
- 빅쿼리의 특징
 - ✓ NoOps, 설치/운영이 필요없다.
 - ✓ SQL 언어 사용
 - ✓ 클라우드 규모의 인프라를 통한 대용량 지원과 빠른 성능
- 빅쿼리의 성능과 스케일은 <https://goo.gl/fDN8YE>의 예를 보면 짐작할 수 있다.



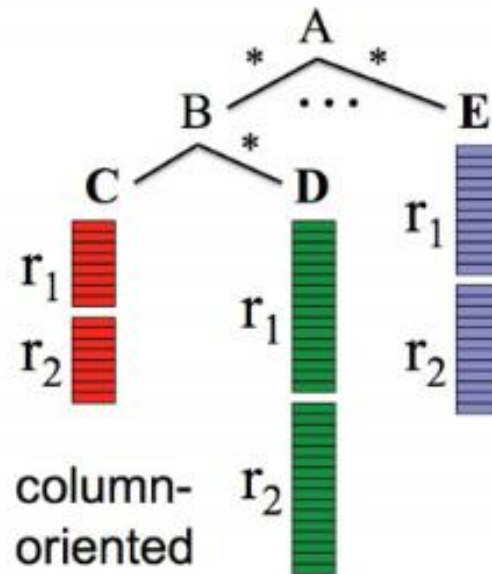
- 그림 11-1 빅쿼리 분석 예 (출처 : <https://goo.gl/fDN8YE>)
- 이 예는 위키피디아에서 1,000억 개의 레코드를 스캔해서 정규표현식으로 "G.*o.*o.*g" 문자열을 찾아내고, 그 문서의 뷰 수를 세고 있다. 대략 4TB 용량의 데이터가 처리되고, 약 30초가 소요된다. 30초 동안, 약 3,300개의 CPU와 330개의 하드디스크와 330Gb의 네트워크가 사용된다. (출처 : <https://goo.gl/ptCDhC>) 이 쿼리를 수행하는 데 소요되는 비용은 딱 \$20다. 일반적인 인프라에서 빅데이터 연산을 하는데 3,300개의 CPU를 동시에 사용하기란 말처럼 쉽지 않은 일이고, 이런 대용량 연산을 단돈 \$20에 처리해주는 것은 대용량 인프라를 공유하는 클라우드 서비스이기에 가능하다.

구글 클라우드의 차별성

- 일반적인 RDBMS : 레코드 단위로 데이터를 저장
- 컬럼 단위로 데이터를 뜯어내어 저장
 - 방대한 양의 데이터를 분석하기 위해 설계



record-oriented



• 빅쿼리, 빅데이터 저장 및 분석 플랫폼

- 빅쿼리의 특징

✓ 데이터 복제를 통한 안정성

- 데이터는 3개로 복제하여 서로 다른 3개의 데이터센터에 분산 저장하기 때문에 유실 위험이 적다.

✓ 배치와 스트리밍 모두 지원

- 한꺼번에 데이터를 로딩하는 일괄 작업(batch) 외에도 REST API 등을 통해서 실시간으로 데이터를 입력할 수 있는 스트리밍 기능을 제공
- 스트리밍 시에는 데이터를 초당 100,000 행씩 입력할 수 있다

✓ 비용 정책

- 클라우드 서비스답게 딱 저장되는 데이터 크기와 쿼리 시에 발생하는 트랜잭션 비용만큼만 과금
- 데이터 저장 요금은 GB당 \$0.02이고, 90일이 지나서 사용하지 않는 데이터는 자동으로 \$0.01로 떨어진다. 일반적으로 가격이 싸다고 알려진 오브젝트 스토리지보다 싸다(구글 클라우드 스토리지는 GB당 \$0.026다).
- 트랜잭션 비용은 쿼리 수행 시 스캔되는 데이터를 기준으로 TB당 \$5다(월 1TB는 무료). 자세한 가격 정책은 <https://cloud.google.com/bigquery/pricing>을 참조

- **쉽다.**

- 하둡이나 스파크로 분석할 때는 맵리듀스 로직을 사용하거나 Spark SQL을 사용하는데, 이 방식은 일정 수준 이상의 전문성이 필요
- 맵리듀스 로직의 경우 전문성 있는 개발자가 분석 로직을 개발해야 하기 때문에 시간이 상대적으로 오래 걸림
- 빅쿼리는 로그인하고 SQL만 수행하면 되니 상대적으로 빅데이터 분석이 쉽다.

- **운영이 필요 없다.**

- 하둡이나 스파크와 같은 빅데이터 솔루션은 설치와 설정, 그리고 클러스터를 유지보수하기가 어려워 별도의 운영 조직이 필요하고 여기에 많은 자원이 소모된다.
- 빅쿼리는 클라우드 서비스기 때문에 개발과 분석에만 집중하면 된다.

- **인프라 투자 없이 막강한 컴퓨팅 자원을 활용한다.**

- 빅쿼리를 이용하면 수천 개의 CPU와 수백/수천 개의 컴퓨팅 자원을 사용할 수 있다.
- 기존 빅데이터 플랫폼도 클라우드 환경에 올리면 가능한 일이지만, 그 설정 노력과 비용 측면에서 차이가 크다.

1. 가입하기

- 구글 클라우드 서비스에 가입하고 로그인

2. Cloud 프로젝트를 선택하거나 만든다



3. 빅쿼리 콘솔로 이동하기

The image consists of two side-by-side screenshots of the Google Cloud Platform (GCP) console. The left screenshot shows the 'Google Cloud Platform' dashboard with a sidebar menu on the left. The 'BigQuery' option is highlighted with a red rectangle. The right screenshot shows the 'BigQuery' console page. The '쿼리 기록' (Query History) tab is selected, and the '쿼리 편집기' (Query Editor) is visible. The page displays a message: '아직 쿼리가 없습니다. 시작하려면 쿼리를 작성하세요.' (No queries yet. Start by writing a query.)

- 공개 데이터세트 쿼리(usa names)

- 미국 이름 데이터 공개 데이터세트를 쿼리해 1910년부터 2013년까지 미국에서 가장 흔한 이름을 확인

The image displays two side-by-side screenshots of the Google Cloud Platform BigQuery console interface.

Left Screenshot: The '데이터 추가' (Add Data) button in the left sidebar is highlighted with a red box. A dropdown menu is open, showing '공개 데이터세트 탐색하기' (Explore public datasets) highlighted with a red box.

Right Screenshot: The search bar at the top of the console is highlighted with a red box and contains the text 'usa names'. The main content area shows the '데이터세트' (Datasets) page with a search result for 'About COVID-19 Public Datasets' and 'Aion On-Chain Transaction Data'.

The screenshot shows the Google Cloud Platform BigQuery console interface. The left sidebar contains navigation links: 쿼리 기록, 저장된 쿼리, 작업 기록, 전송, 예약된 쿼리, 예약, and BI Engine. Below these is a '리소스' section with a '+ 데이터 추가' button and a search bar for '표 및 데이터세트 검색'. The main area displays a SQL query for selecting names and genders from the 'bigquery-public-data.usa_names.usa_1910_2013' dataset, grouped by name and gender, ordered by total count descending, and limited to 10 results. The query is numbered 1 through 11. Below the query editor, there are buttons for '실행' (Execute), '쿼리 저장' (Save query), '보기 저장' (Save view), and '쿼리 예약' (Schedule query). The '실행' button is highlighted with a red box. A status message indicates that the query will process 99.9MB of data. The bottom of the console shows the dataset path 'bigquery-public-data:usa_names'.

USA Names – Marketplace – Da x BigQuery – DataPlatform-01 – C x +

console.cloud.google.c... ☆

Google Cloud Platform DataPlatform-01

BigQuery 기능 및 정보 단축키

쿼리 기록

저장된 쿼리

작업 기록

전송

예약된 쿼리

예약

BI Engine

리소스 + 데이터 추가

표 및 데이터세트 검색

dataplatfrom-01

bigquery-public-data

austin_311

저장되지 않은 쿼리 수정됨 + 전체 화면

```
1 SELECT
2   name, gender,
3   SUM(number) AS total
4 FROM
5   `bigquery-public-data.usa_names.usa_1910_2013`
6 GROUP BY
7   name, gender
8 ORDER BY
9   total DESC
10 LIMIT
11  10
```

SELECT
name, gender,
SUM(number) AS total
FROM
`bigquery-public-
data.usa_names.usa_1910_2013`
GROUP BY
name, gender
ORDER BY
total DESC
LIMIT
10

실행 쿼리 저장 보기 저장 쿼리 예약

더보기

실행 시 이 쿼리가 99.9MB를 처리합니다.

bigquery-public-data:usa_names

빅쿼리 맛보기

BigQuery - DataPlatform-01 - C x +

← → ↺ console.cloud.google.c... ☆ 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99 100 101 102 103 104 105 106 107 108 109 110 111 112 113 114 115 116 117 118 119 120 121 122 123 124 125 126 127 128 129 130 131 132 133 134 135 136 137 138 139 140 141 142 143 144 145 146 147 148 149 150 151 152 153 154 155 156 157 158 159 160 161 162 163 164 165 166 167 168 169 170 171 172 173 174 175 176 177 178 179 180 181 182 183 184 185 186 187 188 189 190 191 192 193 194 195 196 197 198 199 200 201 202 203 204 205 206 207 208 209 210 211 212 213 214 215 216 217 218 219 220 221 222 223 224 225 226 227 228 229 230 231 232 233 234 235 236 237 238 239 240 241 242 243 244 245 246 247 248 249 250 251 252 253 254 255 256 257 258 259 260 261 262 263 264 265 266 267 268 269 270 271 272 273 274 275 276 277 278 279 280 281 282 283 284 285 286 287 288 289 290 291 292 293 294 295 296 297 298 299 300 301 302 303 304 305 306 307 308 309 310 311 312 313 314 315 316 317 318 319 320 321 322 323 324 325 326 327 328 329 330 331 332 333 334 335 336 337 338 339 340 341 342 343 344 345 346 347 348 349 350 351 352 353 354 355 356 357 358 359 360 361 362 363 364 365 366 367 368 369 370 371 372 373 374 375 376 377 378 379 380 381 382 383 384 385 386 387 388 389 390 391 392 393 394 395 396 397 398 399 400 401 402 403 404 405 406 407 408 409 410 411 412 413 414 415 416 417 418 419 420 421 422 423 424 425 426 427 428 429 430 431 432 433 434 435 436 437 438 439 440 441 442 443 444 445 446 447 448 449 450 451 452 453 454 455 456 457 458 459 460 461 462 463 464 465 466 467 468 469 470 471 472 473 474 475 476 477 478 479 480 481 482 483 484 485 486 487 488 489 490 491 492 493 494 495 496 497 498 499 500 501 502 503 504 505 506 507 508 509 510 511 512 513 514 515 516 517 518 519 520 521 522 523 524 525 526 527 528 529 530 531 532 533 534 535 536 537 538 539 540 541 542 543 544 545 546 547 548 549 550 551 552 553 554 555 556 557 558 559 560 561 562 563 564 565 566 567 568 569 570 571 572 573 574 575 576 577 578 579 580 581 582 583 584 585 586 587 588 589 590 591 592 593 594 595 596 597 598 599 600 601 602 603 604 605 606 607 608 609 610 611 612 613 614 615 616 617 618 619 620 621 622 623 624 625 626 627 628 629 630 631 632 633 634 635 636 637 638 639 640 641 642 643 644 645 646 647 648 649 650 651 652 653 654 655 656 657 658 659 660 661 662 663 664 665 666 667 668 669 670 671 672 673 674 675 676 677 678 679 680 681 682 683 684 685 686 687 688 689 690 691 692 693 694 695 696 697 698 699 700 701 702 703 704 705 706 707 708 709 710 711 712 713 714 715 716 717 718 719 720 721 722 723 724 725 726 727 728 729 730 731 732 733 734 735 736 737 738 739 740 741 742 743 744 745 746 747 748 749 750 751 752 753 754 755 756 757 758 759 760 761 762 763 764 765 766 767 768 769 770 771 772 773 774 775 776 777 778 779 780 781 782 783 784 785 786 787 788 789 790 791 792 793 794 795 796 797 798 799 800 801 802 803 804 805 806 807 808 809 810 811 812 813 814 815 816 817 818 819 820 821 822 823 824 825 826 827 828 829 830 831 832 833 834 835 836 837 838 839 840 841 842 843 844 845 846 847 848 849 850 851 852 853 854 855 856 857 858 859 860 861 862 863 864 865 866 867 868 869 870 871 872 873 874 875 876 877 878 879 880 881 882 883 884 885 886 887 888 889 890 891 892 893 894 895 896 897 898 899 900 901 902 903 904 905 906 907 908 909 910 911 912 913 914 915 916 917 918 919 920 921 922 923 924 925 926 927 928 929 930 931 932 933 934 935 936 937 938 939 940 941 942 943 944 945 946 947 948 949 950 951 952 953 954 955 956 957 958 959 960 961 962 963 964 965 966 967 968 969 970 971 972 973 974 975 976 977 978 979 980 981 982 983 984 985 986 987 988 989 990 991 992 993 994 995 996 997 998 999 1000

Google Cloud Platform DataPlatform-01

BigQuery 기능 및 정보 단축키

쿼리 기록

저장된 쿼리

작업 기록

전송

예약된 쿼리

예약

BI Engine

리소스 + 데이터 추가

표 및 데이터세트 검색

dataplatfrom-01

bigquery-public-data

austin_311

austin_bikeshare

austin_crime

austin_incidents

austin_waste

baseball

bitcoin_blockchain

bls

bls_qcew

breathe

쿼리 편집기

새 쿼리 작성

1 SELECT

2 name, gender,

3 SUM(number) AS total

4 FROM

5 `bigquery-public-data.usa_names.usa_1910_2013`

6 GROUP BY

7 name, gender

8 ORDER BY

9 total DESC

10 LIMIT

11 10

실행

쿼리 저장

보기 저장

쿼리 예약

더보기

실행 시 이 쿼리가 99.9MB를 처리합니다.

쿼리 결과

결과 저장

데이터 탐색

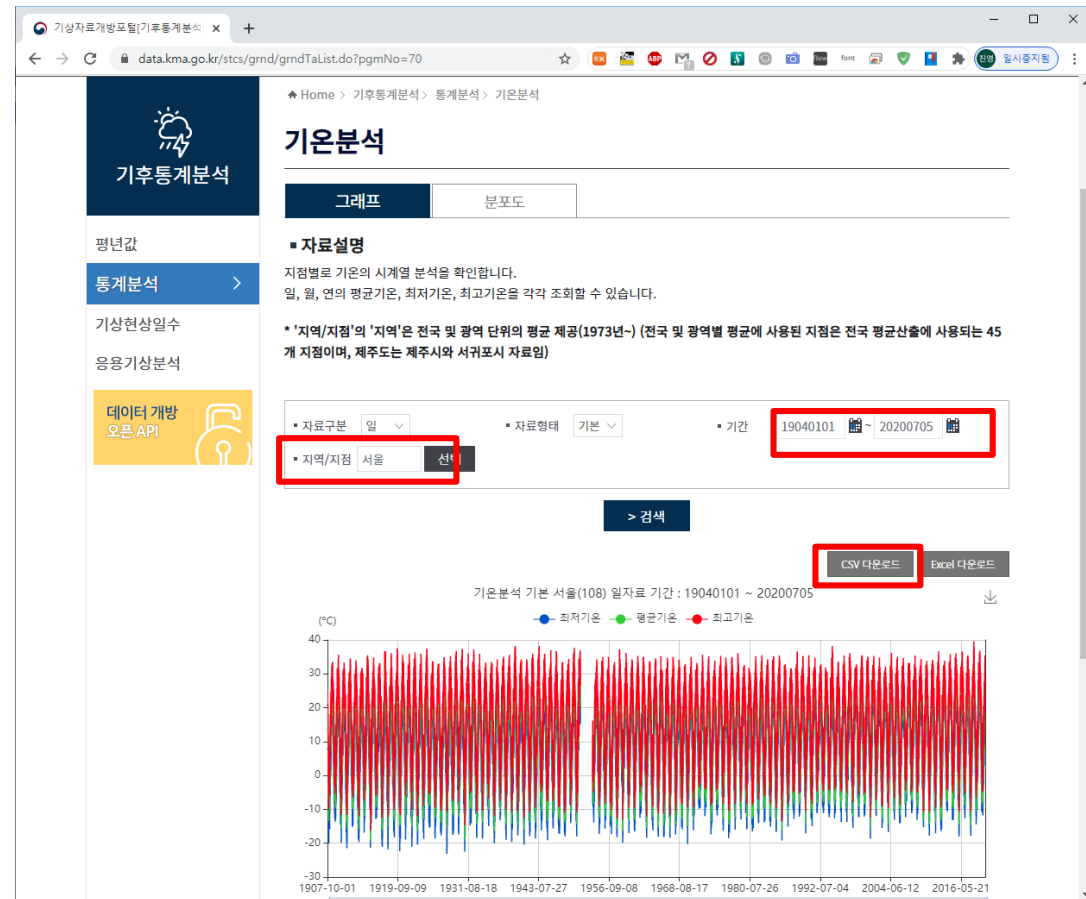
쿼리 완료(2.4초 경과, 99.9MB 처리됨)

작업 정보 결과 JSON 실행 세부정보

행	name	gender	total
1	James	M	4924235
2	John	M	4818746
3	Robert	M	4703680
4	Michael	M	4280040
5	William	M	3811998
6	Mary	F	3728041
7	David	M	3541625
8	Richard	M	2526927

• 데이터 다운로드

- <https://data.kma.go.kr/cmmn/main.do>
- 기후통계분석 → 통계분석 → 기온분석
- 기간 : 1904년 1월1일부터 2020년 7월 6일까지로 설정
- 지역 : 서울
- 검색 후 csv 다운로드



기온분석				
[검색조건]				
자료구분 : 일				
자료형태 : 기본				
지역/지점 : 서울				
기간 : 19040101~20200705				

날짜	지점	평균기온(°C)	최저기온(°C)	최고기온(°C)
#####	108	13.5	7.9	20.7
#####	108	16.2	7.9	22
#####	108	16.2	13.1	21.3
#####	108	16.5	11.2	22
#####	108	17.6	10.9	25.4
#####	108	13	11.2	21.3
#####	108	11.3	6.3	16.1
#####	108	8.9	3.9	14.9
#####	108	11.6	3.8	21.1
#####	108	14.2	6.4	24.1
#####	108	15.4	10.1	20.4
#####	108	13.9	11.1	17.4
#####	108	13.8	8.3	21.3
#####	108	13	6.1	20.6
#####	108	13.1	5.7	20.9
#####	108	14.1	8.2	20.2

date	state	avg	min	max
1907-10-01	108	13.5	7.9	20.7
1907-10-02	108	16.2	7.9	22
1907-10-03	108	16.2	13.1	21.3
1907-10-04	108	16.5	11.2	22
1907-10-05	108	17.6	10.9	25.4
1907-10-06	108	13	11.2	21.3
1907-10-07	108	11.3	6.3	16.1
1907-10-08	108	8.9	3.9	14.9
1907-10-09	108	11.6	3.8	21.1
1907-10-10	108	14.2	6.4	24.1
1907-10-11	108	15.4	10.1	20.4
1907-10-12	108	13.9	11.1	17.4
1907-10-13	108	13.8	8.3	21.3
1907-10-14	108	13	6.1	20.6
1907-10-15	108	13.1	5.7	20.9
1907-10-16	108	14.1	8.2	20.2
1907-10-17	108	16.4	10.3	21.6
1907-10-18	108	14.3	9.8	20.9
1907-10-19	108	13.9	6.7	21.3
1907-10-20	108	18.3	12.4	22.7
1907-10-21	108	15.2	10.7	19.9
1907-10-22	108	15.4	12.1	19.6
1907-10-23	108	13.1	8.1	16.3

The screenshot shows the Google Cloud Platform BigQuery console interface. The browser address bar displays the URL: `console.cloud.google.com/bigquery?hl=ko&_ga=2.158762744.3108...`. The top navigation bar includes the Google Cloud Platform logo, the project name 'DataPlatform-01', and a search bar. The main header shows 'BigQuery' with links for '기능 및 정보' (Features and Information) and '단축키' (Shortcuts).

On the left sidebar, the '쿼리 기록' (Query History) section is expanded, showing a list of queries. The 'datapatform-01' dataset is highlighted in the '리소스' (Resources) section. Below it, the 'bigquery-public-data' dataset is also visible.

The main content area is titled '쿼리 편집기' (Query Editor). It features a toolbar with buttons for '실행' (Execute), '쿼리 저장' (Save Query), '보기 저장' (Save View), '쿼리 예약' (Schedule Query), and '더보기' (More). The dataset name 'datapatform-01' is displayed prominently. A red box highlights the '데이터세트 만들기' (Create Dataset) button. Below the dataset name, a message states: '데이터세트를 사용할 수 없음' (Cannot use dataset), followed by the text: '위의 제어 기능을 사용하여 데이터세트를 생성하고 리소스 트리 구조를 시작하거나 이 프로젝트의 권한을 확인하세요.' (Use the control features above to create the dataset and start the resource tree structure, or check the permissions for this project.)

- 데이터세트ID : weather
- 데이터 위치 : 서울

BigQuery - DataPlatform-01 - C x +

console.cloud.google.com/bigquery?hl=ko&_ga=2.158762744.3108...

Google Cloud Platform DataPlatform-01

BigQuery 기능 및 정보 단축키

쿼리 기록 저장된 쿼리 작업 기록 전송 예약된 쿼리 예약 BI Engine 리소스 + 데이터 추가

표 및 데이터세트 검색

dataplatfrom-01

bigquery-public-data

데이터세트 만들기

데이터세트 ID

weather

데이터 위치 (선택사항)

서울(asia-northeast3)

기본 표 만료

☒ 사용 안함

☐ 테이블 생성 후 경과 일수:

암호화

데이터가 자동으로 암호화됩니다. 암호화 키 관리 솔루션을 선택하세요.

☒ Google 관리 키

구성이 필요하지 않습니다.

☐ 고객 관리 키

Google Cloud Key Management Service를 통해 관리합니다.

데이터세트 만들기 취소

BigQuery – DataPlatform-01 – C x +

console.cloud.google.com/bigquery?hl=ko&_ga=2.1587...

Google Cloud Platform

BigQuery 기능 및 정보

쿼리 기록
저장된 쿼리
작업 기록
전송
예약된 쿼리
예약
BI Engine
리소스 + 데이터 추가

datapatform-01
weather
bigquery-public-data
austin_311
austin_bikeshare
austin_crime

테이블 만들기

소스

다음 항목으로 테이블 만들기: 파일 선택: 파일 형식:

업로드 ta_20200706104320_re.csv 탐색 CSV

대상

☒ 프로젝트 검색 ☐ 프로젝트 이름 입력

프로젝트 이름 데이터세트 이름 테이블 유형

DataPlatform-01 weather 기본 테이블

테이블 이름

weather

스키마

자동 감지
☐ 스키마 및 입력 매개변수

☐ 텍스트로 편집

date:string, state:string, avg:string, min:string, max:string

테이블 만들기 취소

BigQuery - DataPlatform-01 - console.cloud.google.com/bigqu...

Google Cloud Platform DataPlatform-01 제품 및 리소스

BigQuery 기능 및 정보 단축키

쿼리 기록
저장된 쿼리
작업 기록
전송
예약된 쿼리
예약
BI Engine
리소스 + 데이터 추가

표 및 데이터세트 검색

dataplatfrom-01

weather

weather

bigquery-public-data

austin_311
austin_bikeshare
austin_crime
austin_incidents
austin_waste
baseball
bitcoin_blockchain
bls

쿼리 편집기 + 새 쿼리 작성 편집기 숨기기 전체 화면

1 |

저리 위치: asia-northeast3

실행 쿼리 저장 보기 저장 쿼리 예약 더보기

weather

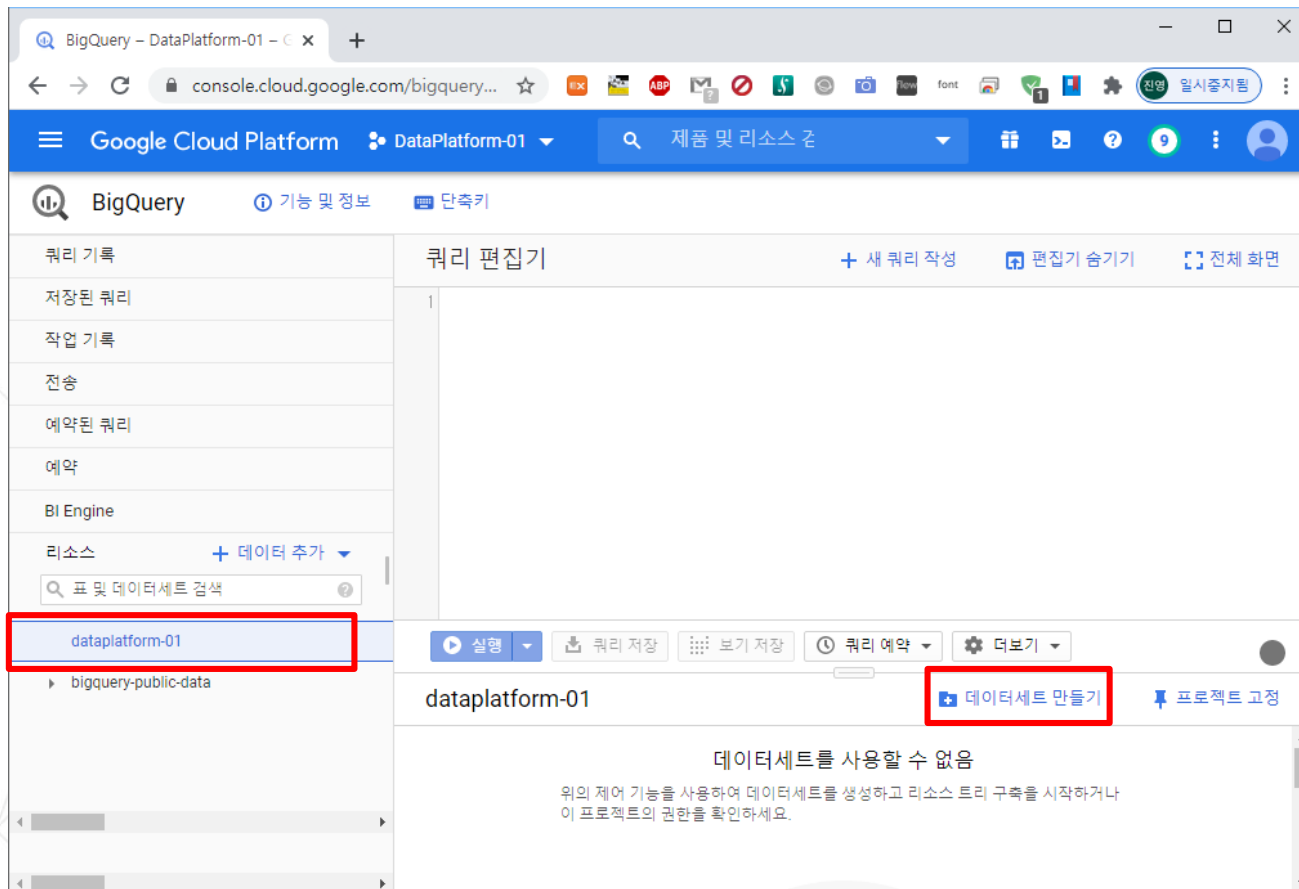
테이블 복사 테이블 삭제 내보내기

스키마 세부정보 미리보기

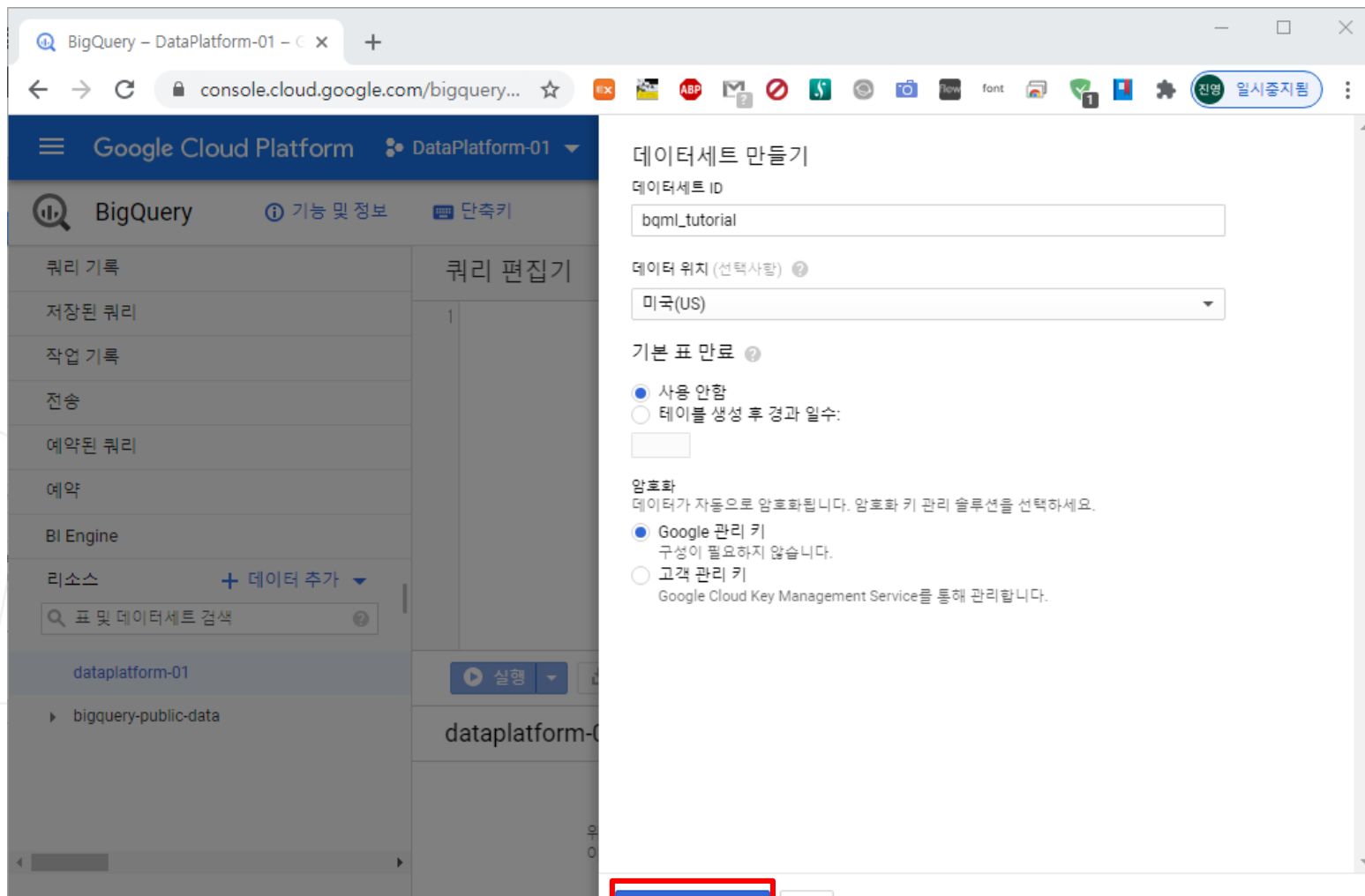
행	date	state	avg	min	max
1	1950-09-01	108	null	null	null
2	1950-09-02	108	null	null	null
3	1950-09-03	108	null	null	null
4	1950-09-04	108	null	null	null
5	1950-09-05	108	null	null	null
6	1950-09-06	108	null	null	null

페이지당 행 수: 100 1 - 100 / 40757 첫 페이지 < > > 마지막 페이지

- Cloud consoli의 프로젝트 선택 페이지에서 cloud 프로젝트를 선택하거나 만든다.
- 데이터 세트 만들기



- 데이터세트 ID : bqml_tutorial
- 데이터 위치 : 미국(US)
 - 현재 공개 데이터세트는 US 멀티 리전 위치에 저장됩니다.



• 모델 만들기

- BigQuery용 출생률 샘플 테이블을 사용하여 선형 회귀 모델을 만든다.

```
#standardSQL
CREATE MODEL `bqml_tutorial.natality_model`
OPTIONS
  (model_type='linear_reg',
   input_label_cols=['weight_pounds']) AS
SELECT
  weight_pounds,
  is_male,
  gestation_weeks,
  mother_age,
  CAST(mother_race AS string) AS mother_race
FROM
  `bigquery-public-data.samples.natality`
WHERE
  weight_pounds IS NOT NULL
  AND RAND() < 0.001
```

- 실행 클릭 후 모델 생성

- CREATE MODEL 문을 사용하여 테이블을 만들므로 쿼리 결과가 표시되지 않는다

쿼리 결과

쿼리 완료(24.5초 경과, 4.1GB(ML) 처리됨)

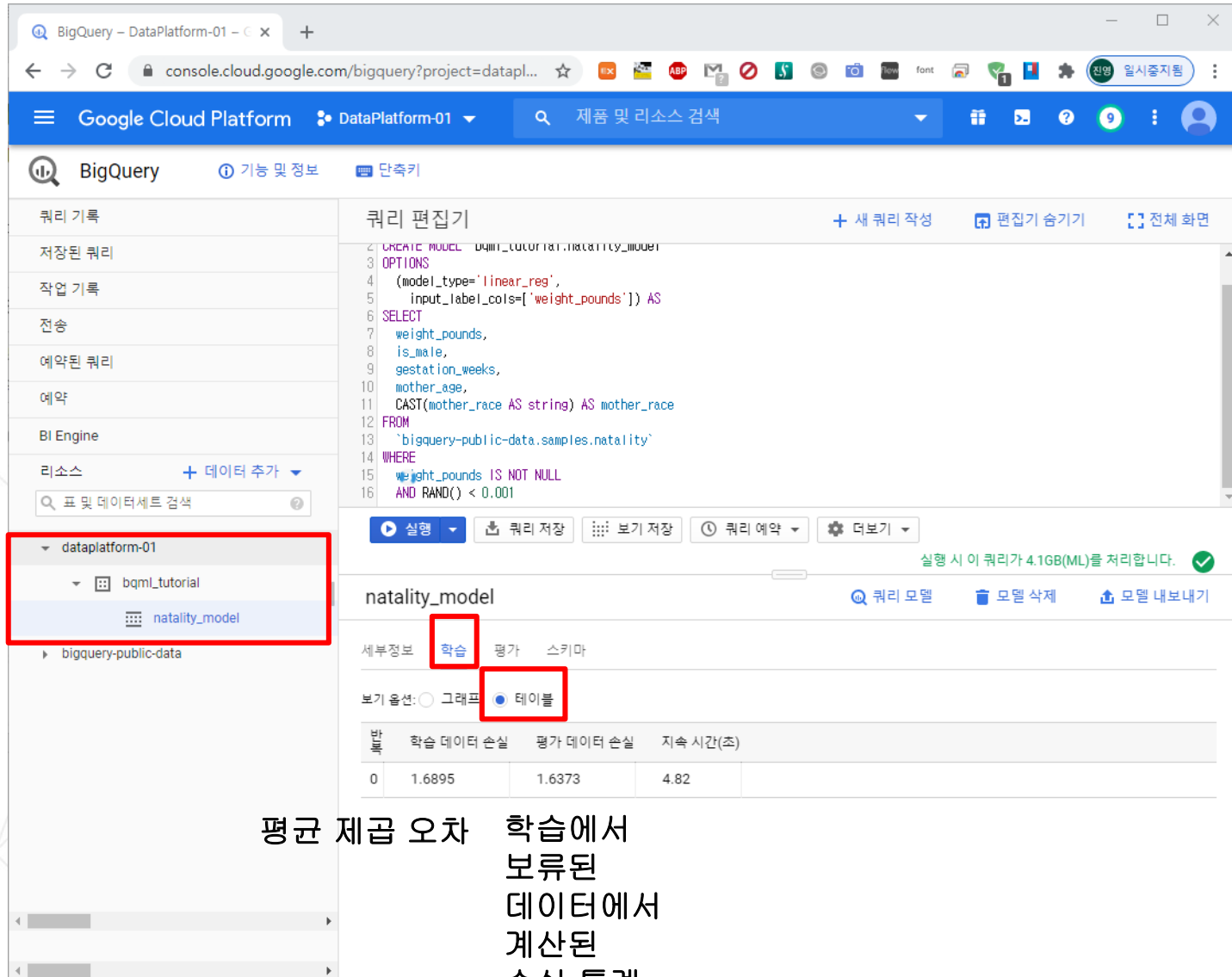
작업 정보 결과 JSON 실행 세부정보

i 이 구문으로 이름이 `datapatform-01:bqml_tutorial.natality_model`인 새 모델이 생성되었습니다.

• 쿼리의 세부정보

- CREATE MODEL → bqml_tutorial.natality_model이라는 모델을 만들고 학습
- OPTIONS(model_type='linear_reg', input_label_cols=['weight_pounds'])
→ 선형 회귀 모델을 만든다
- 선형 회귀는 입력 특성의 선형 조합에서 연속 값을 생성하는 회귀 모델의 한 유형, weight_pounds 열은 입력 라벨 열
- 선형 회귀 모델에서 라벨 열은 실수여야 함
- SELECT 문은 다음 열을 사용하여 아이의 출생 시 체중을 예측
 - ✓ weight_pounds: 아기 체중이며 단위는 파운드 (FLOAT64).
 - ✓ is_male — 남아이면 TRUE, 여아이면 FALSE (BOOL).
 - ✓ gestation_weeks — 임신 주 (INT64).
 - ✓ mother_age — 출산 산모의 나이 (INT64).
 - ✓ mother_race — 산모의 인종에 해당하는 정수 값 (INT64 — 테이블 스키마의 child_race와 같음).
 - 각 고유 값이 서로 다른 카테고리를 나타내며 BigQuery ML에서 mother_race를 숫자가 아닌 특성으로 강제 처리하기 위해 쿼리는 mother_race를 STRING으로 변환
 - 인종은 순서와 척도가 있는 정수보다 카테고리로서 더 많은 의미를 가질 수 있음
- FROM 절인 bigquery-public-data.samples.natality는 샘플 데이터세트에서 출생률 샘플 테이블을 쿼리한다는 것을 나타낸다. 이 데이터세트는 bigquery-public-data 프로젝트에 있다.
- WHERE 절인 WHERE weight_pounds IS NOT NULL AND RAND() < 0.001은 체중이 NULL인 행을 제외하고 RAND 함수를 사용하여 데이터의 샘플을 추출합니다.

- 학습 통계 가져오기



The screenshot shows the Google Cloud Platform BigQuery console. In the left sidebar, the 'bqml_tutorial' dataset is selected, and the 'natality_model' is highlighted. The main panel shows the 'natality_model' details, with the 'Learning Statistics' (학습) tab selected. The 'View as' (보기 옵션) dropdown is set to 'Table' (테이블). The table displays the following statistics:

반복	학습 데이터 손실	평가 데이터 손실	지속 시간(초)
0	1.6895	1.6373	4.82

Below the table, the text '평균 제곱 오차' (Mean Squared Error) is visible, followed by a description: '학습에서 보류된 데이터에서 계산된 손실 통계' (Loss statistics calculated from data held out during training).

• 모델 평가

- ML.EVALUATE 함수를 사용하여 분류 기준의 성능을 평가
- ML.EVALUATE 함수는 실제 데이터를 기준으로 예측 값을 평가

```
#standardSQL
SELECT
  *
FROM
  ML.EVALUATE(MODEL `bqml_tutorial.natality_model`,
    (
      SELECT
        weight_pounds,
        is_male,
        gestation_weeks,
        mother_age,
        CAST(mother_race AS STRING) AS mother_race
      FROM
        `bigquery-public-data.samples.natality`
      WHERE
        weight_pounds IS NOT NULL))
```

• 쿼리의 세부정보

- 맨 위에 있는 SELECT 문은 모델에서 열을 검색
- FROM 절은 bqml_tutorial.natality_model 모델에 ML.EVALUATE 함수를 사용
- 중첩된 SELECT 문과 FROM 절은 CREATE MODEL 쿼리와 동일
- WHERE 절인 WHERE weight_pounds IS NOT NULL은 체중이 NULL인 행을 제외



▶ 실행

📄 쿼리 저장

📊 보기 저장

🕒 쿼리 예약

⚙️ 더보기

실행 시 이 쿼리가 4.1GB를 처리합니다. ✓

쿼리 결과

📄 결과 저장

📊 데이터 탐색

쿼리 완료(5.7초 경과, 4.1GB 처리됨)

작업 정보

결과

JSON실행 세부정보

행	mean_absolute_error	mean_squared_error	mean_squared_log_error	median_absolute_error	r2_score	explained_variance
1	0.9566223947532317	1.675733577942885	0.03424707767813259	0.7380906177195339	0.046450205077517404	0.04645681300008131



- 모델을 사용하여 결과 예측

- 모델을 평가했으므로 다음 단계는 이 모델을 사용하여 결과를 예측
- 모델을 사용하여 와이오밍(WY)에서 출생한 모든 아기의 출생 시 체중을 예측

```
#standardSQL
SELECT
  predicted_weight_pounds
FROM
  ML.PREDICT(MODEL `bqml_tutorial.natality_model`,
    (
      SELECT
        is_male,
        gestation_weeks,
        mother_age,
        CAST(mother_race AS STRING) AS mother_race
      FROM
        `bigquery-public-data.samples.natality`
      WHERE
        state = "WY"))
```

• 쿼리 세부정보

- 최상위 SELECT 문은 predicted_weight_pounds 열을 검색
 - ✓ 이 열은 ML.PREDICT 함수에서 생성
 - ✓ ML.PREDICT 함수를 사용할 때 모델의 출력 열 이름은 predicted_<label_column_name>
 - ✓ 선형 회귀 모델에서 predicted_label은 label의 예상 값
 - ✓ 로지스틱 회귀 모델에서 predicted_label은 두 입력 라벨 중 하나이며 예측 확률이 더 높은 라벨을 따름
- ML.PREDICT 함수는 bqml_tutorial.natality_model 모델을 사용하여 결과를 예측할 때 사용
 - ✓ 이 쿼리의 중첩된 SELECT 문과 FROM 절은 CREATE MODEL 쿼리와 동일
- WHERE 절인 WHERE state = "WY"는 예측이 와이오밍 주에 한정되어 있음을 나타냅니다.

행	predicted_weight_pounds
1	7.206666513688106
2	8.17438199854223
3	7.531909933137285
4	7.578702417031309
5	7.628818374902039
6	7.423698978287575