DECLARE date\_start, date\_end DATE;

SET date\_end = (SELECT DATE(TIMESTAMP\_MILLIS($start\_time$)));

SET date\_start = DATE\_SUB(date\_end, INTERVAL 30 DAY);

CREATE OR REPLACE TABLE `twttr-recos-ml-prod.realgraph.candidates\_for\_training`

PARTITION BY ds

AS

WITH T1 AS (

SELECT source\_id, destination\_id, label, dateHour

FROM `twttr-bq-cassowary-prod.user.interaction\_graph\_labels\_daily`

LEFT JOIN UNNEST(labels) AS label

WHERE DATE(dateHour) BETWEEN date\_start AND date\_end

), T2 AS (

SELECT source\_id, destination\_id, num\_tweets

FROM `twttr-recos-ml-prod.realgraph.tweeting\_follows`

), T3 AS (

SELECT

COALESCE(T1.source\_id, T2.source\_id) AS source\_id,

COALESCE(T1.destination\_id, T2.destination\_id) AS destination\_id,

COUNT(DISTINCT(T1.dateHour)) AS num\_days,

MIN(COALESCE(num\_tweets,0)) AS num\_tweets, -- all rows' num\_tweets should be the same

COALESCE(DATE\_DIFF(date\_end, DATE(MAX(T1.dateHour)), DAY),30) AS days\_since\_last\_interaction,

COUNT(DISTINCT(label)) AS label\_types,

COUNTIF(label="num\_follows") AS num\_follows,

COUNTIF(label="num\_favorites") AS num\_favorites,

COUNTIF(label="num\_tweet\_clicks") AS num\_tweet\_clicks,

COUNTIF(label="num\_profile\_views") AS num\_profile\_views,

FROM T1

FULL JOIN T2

USING (source\_id, destination\_id)

LEFT JOIN `twttr-bq-cassowary-prod.user.interaction\_graph\_agg\_negative\_edge\_snapshot` N

USING (source\_id, destination\_id)

WHERE N.source\_id IS NULL AND N.destination\_id IS NULL

GROUP BY 1,2

ORDER BY 3 DESC,4 DESC

), T4 AS (

SELECT RANK() OVER (PARTITION BY source\_id ORDER BY num\_days DESC, num\_tweets DESC) AS rn, \*

FROM T3

) SELECT \*, date\_end AS ds FROM T4 WHERE rn <= 2000;

SELECT ds FROM `twttr-recos-ml-prod.realgraph.candidates\_for\_training`

WHERE ds = (SELECT DATE(TIMESTAMP\_MILLIS($start\_time$)))

LIMIT 1