# Create table:

CREATE TABLE Badges (

Id int(10) NOT NULL,

UserId int(10) NOT NULL,

Name nvarchar(50) NOT NULL,

Date datetime NOT NULL,

Class tinyint(3) NOT NULL,

TagBased bit NOT NULL,

PRIMARY KEY ( Id )

);

CREATE TABLE Comments (

Id int(10) NOT NULL,

PostId int(10) NOT NULL,

Score int(10) NOT NULL,

Text nvarchar(600) NOT NULL,

CreationDate datetime NOT NULL,

UserDisplayName nvarchar(30) NULL,

UserId int(10) NULL,

PRIMARY KEY ( Id )

);

CREATE TABLE PostHistory (

Id int(10) NOT NULL,

PostHistoryTypeId tinyint(3) NOT NULL,

PostId int(10) NOT NULL,

RevisionGUID nvarchar(800) NOT NULL,

CreationDate datetime NOT NULL,

UserId int(10) NULL,

UserDisplayName nvarchar(40) NULL,

Comment nvarchar(400) NULL,

Text nvarchar(800) NULL,

PRIMARY KEY ( Id )

);

CREATE TABLE PostLinks (

Id int(10) NOT NULL,

CreationDate datetime NOT NULL,

PostId int(10) NOT NULL,

RelatedPostId int(10) NOT NULL,

LinkTypeId tinyint(3) NOT NULL,

PRIMARY KEY ( Id )

);

CREATE TABLE Posts (

Id int(10) NOT NULL,

PostTypeId tinyint(3) NOT NULL,

AcceptedAnswerId int(10) NULL,

ParentId int(10) NULL,

CreationDate datetime NOT NULL,

DeletionDate datetime NULL,

Score int(10) NOT NULL,

ViewCount int(10) NULL,

Body nvarchar(8000) NULL,

OwnerUserId int(10) NULL,

OwnerDisplayName nvarchar(40) NULL,

LastEditorUserId int(10) NULL,

LastEditorDisplayName nvarchar(40) NULL,

LastEditDate datetime NULL,

LastActivityDate datetime NULL,

Title nvarchar(250) NULL,

Tags nvarchar(250) NULL,

AnswerCount int(10) NULL,

CommentCount int(10) NULL,

FavoriteCount int(10) NULL,

ClosedDate datetime NULL,

CommunityOwnedDate datetime NULL,

PRIMARY KEY ( Id )

);

CREATE TABLE Tags (

Id int(10) NOT NULL,

TagName nvarchar(35) NULL,

Count int(10) NOT NULL,

ExcerptPostId int(10) NULL,

WikiPostId int(10) NULL,

PRIMARY KEY ( Id )

);

CREATE TABLE Users (

Id int(10) NOT NULL,

Reputation int(10) NOT NULL,

CreationDate datetime NOT NULL,

DisplayName nvarchar(40) NULL,

LastAccessDate datetime NOT NULL,

WebsiteUrl nvarchar(200) NULL,

Location nvarchar(100) NULL,

AboutMe nvarchar(800) NULL,

Views int(10) NOT NULL,

UpVotes int(10) NOT NULL,

DownVotes int(10) NOT NULL,

ProfileImageUrl nvarchar(200) NULL,

EmailHash varchar(32) NULL,

AccountId int(10) NULL,

PRIMARY KEY ( Id )

);

CREATE TABLE Votes (

Id int(10) NOT NULL,

PostId int(10) NOT NULL,

VoteTypeId tinyint(3) NOT NULL,

UserId int(10) NULL,

CreationDate datetime NULL,

BountyAmount int(10) NULL,

PRIMARY KEY ( Id )

);

# Load data

set global local\_infile = 1;

LOAD XML LOCAL INFILE 'D:\\MySQL\\MySQL Server 8.0\\Uploads\\posts.xml'

INTO TABLE posts

ROWS IDENTIFIED BY '<row>';

LOAD XML LOCAL INFILE 'D:\\MySQL\\MySQL Server 8.0\\Uploads\\Badges.xml'

INTO TABLE Badges

ROWS IDENTIFIED BY '<row>';

LOAD XML LOCAL INFILE 'D:\\MySQL\\MySQL Server 8.0\\Uploads\\Comments.xml'

INTO TABLE Comments

ROWS IDENTIFIED BY '<row>';

LOAD XML LOCAL INFILE 'D:\\MySQL\\MySQL Server 8.0\\Uploads\\PostHistory.xml'

INTO TABLE PostHistory

ROWS IDENTIFIED BY '<row>';

LOAD XML LOCAL INFILE 'D:\\MySQL\\MySQL Server 8.0\\Uploads\\postlinks.xml'

INTO TABLE postlinks

ROWS IDENTIFIED BY '<row>';

LOAD XML LOCAL INFILE 'D:\\MySQL\\MySQL Server 8.0\\Uploads\\tags.xml'

INTO TABLE tags

ROWS IDENTIFIED BY '<row>';

LOAD XML LOCAL INFILE 'D:\\MySQL\\MySQL Server 8.0\\Uploads\\users.xml'

INTO TABLE users

ROWS IDENTIFIED BY '<row>';

LOAD XML LOCAL INFILE 'D:\\MySQL\\MySQL Server 8.0\\Uploads\\votes.xml'

INTO TABLE votes

ROWS IDENTIFIED BY '<row>';

# Data Analysis

CREATE table v202303\_posts\_after\_20201201 as

SELECT \* from stackoverflow202303.posts

where stackoverflow202303.posts.CreationDate >= '2020-12-01';

CREATE TABLE v202303\_posts\_after\_20201201\_q\_to\_a AS

SELECT stackoverflow202303.posts.\* FROM stackoverflow202303.posts

INNER JOIN v202303\_posts\_after\_20201201 ON v202303\_posts\_after\_20201201.ParentId = stackoverflow202303.posts.Id;

CREATE table v202403\_posts\_after\_20211201 as

SELECT \* from stackoverflow202403.posts

where stackoverflow202403.posts.CreationDate >= '2021-12-01';

CREATE TABLE v202403\_posts\_after\_20211201\_q\_to\_a AS

SELECT stackoverflow202403.posts.\* FROM stackoverflow202403.posts

INNER JOIN v202403\_posts\_after\_20211201 ON v202403\_posts\_after\_20211201.ParentId = stackoverflow202403.posts.Id;

# v2022-users

CREATE table users\_in\_v202303 as

select stackoverflow202303.users.\*

from stackoverflow202303.users

INNER JOIN

(SELECT OwnerUserId FROM v202303\_posts\_after\_20201201

UNION

SELECT OwnerUserId FROM v202303\_posts\_after\_20201201\_q\_to\_a

) u on u.OwnerUserId=stackoverflow202303.users.Id;

# v2023-users

CREATE table users\_in\_v202403 as

select stackoverflow202403.users.\*

from stackoverflow202403.users

INNER JOIN

(SELECT OwnerUserId FROM v202403\_posts\_after\_20211201

UNION

SELECT OwnerUserId FROM v202403\_posts\_after\_20211201\_q\_to\_a

) u on u.OwnerUserId=stackoverflow202403.users.Id;

# post vote within7days

create table v202403\_vote\_within7days as

(SELECT p.Id,

SUM(CASE WHEN v.VoteTypeId = 1 THEN 1 ELSE 0 END) AS VoteType1Count,

SUM(CASE WHEN v.VoteTypeId = 2 THEN 1 ELSE 0 END) AS VoteType2Count,

SUM(CASE WHEN v.VoteTypeId = 3 THEN 1 ELSE 0 END) AS VoteType3Count

FROM v202403\_posts\_after\_20211201 AS p

JOIN stackoverflow202403.votes AS v ON p.Id = v.PostId

WHERE

-- p.PostTypeId = 2

-- AND

p.CreationDate BETWEEN DATE\_SUB(p.CreationDate, INTERVAL 7 DAY) AND p.CreationDate

GROUP BY p.Id)

# post vote within7days

create table v202303\_vote\_within7days as

(SELECT p.Id,

SUM(CASE WHEN v.VoteTypeId = 1 THEN 1 ELSE 0 END) AS VoteType1Count,

SUM(CASE WHEN v.VoteTypeId = 2 THEN 1 ELSE 0 END) AS VoteType2Count,

SUM(CASE WHEN v.VoteTypeId = 3 THEN 1 ELSE 0 END) AS VoteType3Count

FROM v202303\_posts\_after\_20201201 AS p

JOIN stackoverflow202303.votes AS v ON p.Id = v.PostId

WHERE

-- p.PostTypeId = 2

-- AND

p.CreationDate BETWEEN DATE\_SUB(p.CreationDate, INTERVAL 7 DAY) AND p.CreationDate

GROUP BY p.Id)

CREATE TABLE users\_information\_v202403 AS

SELECT

ui.Id,

TIMESTAMPDIFF(YEAR, ui.CreationDate, '2021-12-01') AS UsersAge,

ui.Reputation AS UsersReputation,

(SELECT COUNT(\*) FROM stackoverflow202403.posts WHERE PostTypeId = 1 AND OwnerUserId = ui.Id AND CreationDate < '2021-12-01') AS PostType1Count,

(SELECT COUNT(\*) FROM stackoverflow202403.posts WHERE PostTypeId = 1 AND OwnerUserId = ui.Id AND CreationDate BETWEEN '2020-12-01' AND '2021-12-01') AS PostType1CountLastYear,

(SELECT COUNT(\*) FROM stackoverflow202403.posts WHERE PostTypeId = 1 AND OwnerUserId = ui.Id AND CreationDate BETWEEN '2018-12-01' AND '2021-12-01') AS PostType1CountThreeYear,

(SELECT COUNT(\*) FROM stackoverflow202403.posts WHERE PostTypeId = 2 AND OwnerUserId = ui.Id AND CreationDate < '2021-12-01') AS PostType2Count,

(SELECT COUNT(\*) FROM stackoverflow202403.posts WHERE PostTypeId = 2 AND OwnerUserId = ui.Id AND CreationDate BETWEEN '2020-12-01' AND '2021-12-01') AS PostType2CountLastYear,

(SELECT COUNT(\*) FROM stackoverflow202403.posts WHERE PostTypeId = 2 AND OwnerUserId = ui.Id AND CreationDate BETWEEN '2018-12-01' AND '2021-12-01') AS PostType2CountThreeYear,

(SELECT COUNT(\*) FROM stackoverflow202403.posts WHERE OwnerUserId = ui.Id AND CreationDate < '2021-12-01') AS PostCount,

(SELECT COUNT(\*) FROM stackoverflow202403.posts WHERE OwnerUserId = ui.Id AND CreationDate BETWEEN '2018-12-01' AND '2021-12-01') AS PostCountThreeYear,

(SELECT COUNT(\*) FROM stackoverflow202403.posts WHERE OwnerUserId = ui.Id AND CreationDate BETWEEN '2020-12-01' AND '2021-12-01') AS PostCountLastYear,

(SELECT COUNT(\*) FROM stackoverflow202403.comments WHERE UserId = ui.Id AND CreationDate < '2021-12-01') AS CommentCount,

(SELECT COUNT(\*) FROM stackoverflow202403.comments WHERE UserId = ui.Id AND CreationDate BETWEEN '2018-12-01' AND '2021-12-01') AS CommentCountThreeYear,

(SELECT COUNT(\*) FROM stackoverflow202403.comments WHERE UserId = ui.Id AND CreationDate BETWEEN '2020-12-01' AND '2021-12-01') AS CommentCountLastYear

FROM users\_in\_v202403 ui;

CREATE TABLE users\_information\_v202303 AS

SELECT

ui.Id,

TIMESTAMPDIFF(YEAR, ui.CreationDate, '2020-12-01') AS UsersAge,

ui.Reputation AS UsersReputation,

(SELECT COUNT(\*) FROM stackoverflow202403.posts WHERE PostTypeId = 1 AND OwnerUserId = ui.Id AND CreationDate < '2020-12-01') AS PostType1Count,

(SELECT COUNT(\*) FROM stackoverflow202403.posts WHERE PostTypeId = 1 AND OwnerUserId = ui.Id AND CreationDate BETWEEN '2019-12-01' AND '2020-12-01') AS PostType1CountLastYear,

(SELECT COUNT(\*) FROM stackoverflow202403.posts WHERE PostTypeId = 1 AND OwnerUserId = ui.Id AND CreationDate BETWEEN '2017-12-01' AND '2020-12-01') AS PostType1CountThreeYear,

(SELECT COUNT(\*) FROM stackoverflow202403.posts WHERE PostTypeId = 2 AND OwnerUserId = ui.Id AND CreationDate < '2020-12-01') AS PostType2Count,

(SELECT COUNT(\*) FROM stackoverflow202403.posts WHERE PostTypeId = 2 AND OwnerUserId = ui.Id AND CreationDate BETWEEN '2019-12-01' AND '2020-12-01') AS PostType2CountLastYear,

(SELECT COUNT(\*) FROM stackoverflow202403.posts WHERE PostTypeId = 2 AND OwnerUserId = ui.Id AND CreationDate BETWEEN '2017-12-01' AND '2020-12-01') AS PostType2CountThreeYear,

(SELECT COUNT(\*) FROM stackoverflow202403.posts WHERE OwnerUserId = ui.Id AND CreationDate < '2020-12-01') AS PostCount,

(SELECT COUNT(\*) FROM stackoverflow202403.posts WHERE OwnerUserId = ui.Id AND CreationDate BETWEEN '2017-12-01' AND '2020-12-01') AS PostCountThreeYear,

(SELECT COUNT(\*) FROM stackoverflow202403.posts WHERE OwnerUserId = ui.Id AND CreationDate BETWEEN '2019-12-01' AND '2020-12-01') AS PostCountLastYear,

(SELECT COUNT(\*) FROM stackoverflow202403.comments WHERE UserId = ui.Id AND CreationDate < '2020-12-01') AS CommentCount,

(SELECT COUNT(\*) FROM stackoverflow202403.comments WHERE UserId = ui.Id AND CreationDate BETWEEN '2017-12-01' AND '2020-12-01') AS CommentCountThreeYear,

(SELECT COUNT(\*) FROM stackoverflow202403.comments WHERE UserId = ui.Id AND CreationDate BETWEEN '2019-12-01' AND '2020-12-01') AS CommentCountLastYear

FROM users\_in\_v202303 ui;

-- comments

create table v202303\_comment\_after\_20201201 as

select

t.Id,

t.OwnerUserId,

t.PostTypeId,

t.CreationDate AS PostCreationDate,

t2.CreationDate,

t2.PostId,

t2.Id as CommentId,

t2.UserId,

t2.Score,

t2.CreationDate as CommentCreationDate,

t2.Text

from v202303\_posts\_after\_20201201 t

inner join stackoverflow202303.comments t2 on t2.PostId=t.Id;

create table v202403\_comment\_after\_20211201 as

select

t.Id,

t.OwnerUserId,

t.PostTypeId,

t.CreationDate AS PostCreationDate,

t2.CreationDate,

t2.PostId,

t2.Id as CommentId,

t2.UserId,

t2.Score,

t2.CreationDate as CommentCreationDate,

t2.Text

from v202403\_posts\_after\_20211201 t

inner join stackoverflow202403.comments t2 on t2.PostId=t.Id;