User Behavior Analysis on Collaborative Q&A Platform - StackOverflow

**Data Gathering**

* From the above link we have downloaded StackOverflow data

<http://data.stackexchange.com/stackoverflow/query/new>

**Following is the list of queries we have used to download data**

########################## Questions #########################################

SELECT p.id as "QueID",

p.AcceptedANswerID as "QueAcceptedAnswerId",

p.Creationdate as "QueCreationDate",

p.ClosedDate as "QueClosedDate",

p.DeletionDate as "QueDeletionDate",

p.Score as "QueScore",

p.viewcount as "QueviewCount",

p.body as "QueBody",

p.tags as "QueTags",

p.Answercount as "QueAnswerCount",

p.commentcount as "QueCommnetCount",

p.favoritecount as "QueFavoriteCount",

u.id as "QueOwnerUserID",

u.reputation as "QueOwnerReputation"

FROM posts p , users u

WHERE p.owneruserid= u.id and p.PostTypeId=1

AND p.CreationDate BETWEEN '2012-03-01 00:00:00' AND '2012-03-15 23:59:59'

AND p.Answercount >1

Data File Name: **Qapr 1-5.csv** [Questions 1st April – 15th April]

################################# Answers #####################################

SELECT p.id as "AnsID",

p.ParentId as "AnsParentId",

p.Creationdate as "AnsCreationdate",

p.DeletionDate as "AnsDeletionDate",

p.ClosedDate as "AnsClosedDate",

p.Score as "AnsScore",

p.viewcount as "AnsViewCount",

p.body as "AnsBody",

p.tags as "AnsTags",

p.commentcount as "AnsCommnetCount",

p.favoritecount as "AnsFavoriteCount",

u.id as "AnsUserID",

u.reputation as "AnsReputation"

FROM posts p , users u

WHERE p.owneruserid= u.id and p.PostTypeId=2 and p.parentid in (

SELECT pq.id

FROM posts pq , users u

WHERE pq.owneruserid= u.id and pq.PostTypeId=1

AND pq.CreationDate BETWEEN '2012-03-01 00:00:00' AND '2012-03-05 23:59:59'

AND pq.Answercount >1)

Data File Name: **Aapr 1-5.csv** [Answers 1st April – 15th April]

################### Answers Votes #######################################

SELECT v.postid "AnsID",v.creationdate "VoteCreationDate",

(case when v.VoteTypeId =1 then 1 else 0 end) as "AnsAccepted",

(case when v.VoteTypeId =2 then 1 else 0 end) as "AnsUpvotecount",

(case when v.VoteTypeId =3 then 1 else 0 end) as "AnsDownvotecount",

(case when v.VoteTypeId =9 then v.BountyAmount else 0 end) as "AnsBounty"

FROM Votes V

WHERE postid in (SELECT p.id

FROM posts p , users u

WHERE p.owneruserid= u.id and p.PostTypeId=2 and p.parentid in (

SELECT pq.id

FROM posts pq , users u

WHERE pq.owneruserid= u.id and pq.PostTypeId=1

AND pq.CreationDate BETWEEN '2012-03-01 00:00:00' AND '2012-03-02 23:59:59'

AND pq.Answercount >1))

Order by v.postid, v.creationdate

Data File Name: **AnsVoteApr1-2.csv** [Answer Votes 1st April –2nd April]

################### Answers Reputation #######################################

We have calculated reputation at the time when users were answering questions

For Example,

* Calculated user reputation on 1st April and using it as their reputation for answers posted between 1st April and 15th April
* Again calculating reputation for all users on 15th April and using it as their reputation for answers posted between 16th April and 30th April

SELECT p.OwnerUserId UserId,

SUM(

CASE

WHEN v.VoteTypeId = 1

THEN 15

WHEN v.VoteTypeId = 2 AND p.PostTypeId = 1

THEN 5

WHEN v.VoteTypeId = 2 AND p.PostTypeId = 2

THEN 10

WHEN v.VoteTypeId = 3

THEN -2

WHEN v.VoteTypeId = 8

THEN -1\*(v.BountyAmount)

WHEN v.VoteTypeId = 9 AND v.BountyAmount IS NOT NULL

THEN v.BountyAmount

ELSE 0

END) AReputation

FROM Votes v

INNER JOIN POSTS p

ON p.Id= v.PostId

INNER JOIN users u

ON u.id= p.owneruserid

WHERE p.posttypeid IN (1,2)

AND (p.creationdate < '2012-04-01' )

AND u.id IN

(SELECT u.id

FROM posts p ,

users u

WHERE p.owneruserid= u.id

AND p.PostTypeId =2

AND p.parentid IN

(SELECT pq.id

FROM posts pq ,

users u

WHERE pq.owneruserid= u.id

AND pq.PostTypeId =1

AND pq.CreationDate BETWEEN '2012-04-01 00:00:00' AND '2012-04-15 23:59:59'

)

)

GROUP BY p.OwnerUserId;

**Data File Name: ARep\_Apr 1-15.csv [Reputation of answerers at time of posting answers between 1stApril –15th April]**

####################### Question Reputation ###################3

We have calculated reputation at the time when users were posting questions

For Example,

* Calculated user reputation on 1st April and using it as their reputation for questions posted between 1st April and 15th April
* Again calculating reputation for all users on 15th April and using it as their reputation for questions posted between 16th April and 30th April

SELECT p.OwnerUserId UserId,

SUM(

CASE

WHEN v.VoteTypeId = 1

THEN 15

WHEN v.VoteTypeId = 2

AND p.PostTypeId = 1

THEN 5

WHEN v.VoteTypeId = 2

AND p.PostTypeId = 2

THEN 10

WHEN v.VoteTypeId = 3

THEN -2

WHEN v.VoteTypeId = 8

THEN -1\*(v.BountyAmount)

WHEN v.VoteTypeId = 9

AND v.BountyAmount IS NOT NULL

THEN v.BountyAmount

ELSE 0

END) QReputation

FROM Votes v

INNER JOIN POSTS p

ON p.Id= v.PostId

INNER JOIN users u

ON u.id = p.owneruserid

WHERE p.posttypeid IN (1,2)

AND (p.creationdate < '2012-03-01')

AND u.id IN

(SELECT u.id

FROM posts p ,

users u

WHERE p.owneruserid= u.id

AND p.PostTypeId =1

AND p.CreationDate BETWEEN '2012-03-01 00:00:00' AND '2012-03-15 23:59:59'

AND p.Answercount >1

)

GROUP BY p.OwnerUserId;

**Data File Name: QRep\_Apr 1-15.csv [Reputation of askers at time of posting answers between 1stApril –15th April]**

################ Most popular StackOverflow tags in any given month #################

select

num.TagName as Tag,

row\_number() over (order by rate.Rate desc) as MayRank,

row\_number() over (order by num.Num desc) as TotalRank,

rate.Rate as QuestionsInMay,

num.Num as QuestionsTotal

from

(select count(PostId) as Rate, TagName

from

Tags, PostTags, Posts

where Tags.Id = PostTags.TagId and Posts.Id = PostId

and Posts.CreationDate < '2012-04-01'

and Posts.CreationDate > '2012-03-01'

group by TagName) as rate

INNER JOIN

(select count(PostId) as Num, TagName

from

Tags, PostTags, Posts

where Tags.Id = PostTags.TagId and Posts.Id = PostId

group by TagName

having count(PostId) > 800)

as num ON rate.TagName = num.TagName

order by rate.rate desc

#########################Tag Specific Reputation ###########################

Calculated tag specific reputation for 5 most popular tags during May 2012 to March 2012

SELECT p.OwnerUserId UserId,

SUM(

CASE

WHEN v.VoteTypeId = 1

THEN 15

WHEN v.VoteTypeId = 2 AND p.PostTypeId = 1

THEN 5

WHEN v.VoteTypeId = 2 AND p.PostTypeId = 2

THEN 10

WHEN v.VoteTypeId = 3

THEN -2

WHEN v.VoteTypeId = 8

THEN -1\*(v.BountyAmount)

WHEN v.VoteTypeId = 9 AND v.BountyAmount IS NOT NULL

THEN v.BountyAmount

ELSE 0

END) AReputation

FROM Votes v

INNER JOIN POSTS p

ON p.Id= v.PostId

INNER JOIN users u

ON u.id= p.owneruserid

WHERE p.posttypeid IN (1,2)

AND (p.creationdate < '2012-05-16' )

AND u.id IN

(SELECT u.id

FROM posts p

INNER JOIN users u

ON p.owneruserid= u.id

WHERE p.PostTypeId =2

AND p.parentid IN

(SELECT pq.id

FROM posts pq

INNER JOIN users u

ON pq.owneruserid= u.id

INNER JOIN PostTags pt

ON pt.PostID = pq.Id

INNER JOIN tags t

ON t.Id = pt.TagId

WHERE pq.PostTypeId =1

AND pq.CreationDate BETWEEN '2012-05-16 00:00:00' AND '2012-05-31 23:59:59'

AND t.TagName ='android'

)

)

GROUP BY p.OwnerUserId

ORDER BY p.OwnerUserId

**Data File Name: ARep\_Apr\_php 1-15.csv [Php Tag Specific Reputation of answerers at time of posting answers between 1stApril –15th April]**

#########################Tag Specific Badge Count #######################

Calculated Gold, Silver and Bronze tag specific badge status for all users (1 if badge is given to a user else 0)

SELECT p.OwnerUserId AS UserID,

(CASE WHEN (SUM(CASE

WHEN b.Class = 1

THEN 1

ELSE 0

END))>0 THEN 1

ELSE 0

END) AS GoldBadge,

(CASE WHEN (SUM(CASE

WHEN b.Class = 2

THEN 1

ELSE 0

END)) >0 THEN 1

ELSE 0

END) AS SilverBadge,

(CASE WHEN (SUM(CASE

WHEN b.Class = 3

THEN 1

ELSE 0

END)) >0 THEN 1

ELSE 0

END) AS BronzeBadge

FROM POSTS p

INNER JOIN Badges b

ON b.UserId = p.OwnerUserId

WHERE p.posttypeid IN (1,2)

AND p.OwnerUserId IN

(SELECT u.id

FROM posts p

INNER JOIN users u

ON p.owneruserid = u.id

WHERE p.PostTypeId =2

AND p.parentid IN

(SELECT pq.id

FROM posts pq

INNER JOIN users u

ON pq.owneruserid= u.id

INNER JOIN PostTags pt

ON pt.PostID = pq.Id

INNER JOIN tags t

ON t.Id = pt.TagId

WHERE pq.PostTypeId =1

AND pq.CreationDate BETWEEN '2012-03-16 00:00:00' AND '2012-03-31 23:59:59'

AND t.TagName ='android'

)

)

AND b.TagBased=1

AND b.Name = 'android'

AND b.date < '2012-03-16'

GROUP BY p.OwnerUserId

ORDER BY p.OwnerUserId

**Data File Name: ABadge\_Apr\_android 1-15.csv [Android Tag Specific Badge Status of answerers at time of posting answers between 1stApril –15th April]**