SECTION 1: DATA LOADING AND SEPARATION (using AWS- EMR and AWS - S3):

- 1) Given the 2 files.
 - Events.csv
 - Users.csv
- 2) I noticed that the files files were relatively large and I thought that in the practical scenario the files would be much larger.
 - 3) I set up a 3 node hadoop cluster on AWS to group and analyse the data.

```
I uploaded the files to Amazon S3 and created external tables to reference the files
drop table whisper events;
CREATE EXTERNAL TABLE whisper events(test cohort STRING, event name STRING, user id
STRING, whisper id STRING, extra information STRING, time generated STRING)
ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
STORED AS TEXTFILE LOCATION 's3n://whisper test/events';
msck repair table whisper_events;
drop table whisper users;
CREATE EXTERNAL TABLE whisper_users(user_id STRING, dtype_appversion STRING, ts_created
STRING)
ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
STORED AS TEXTFILE LOCATION 's3n://whisper test/users';
msck repair table whisper users;
   4) I guickly ran a group by clause on the events.csv file on test cohort and identified 6 test
```

groups. [A ... F]

Select test_cohort from whisper_events group by test_cohort;

5) The next step would be to perform an inner join on users.csv file with the events.csv file on user id for each separate test cohort and write the results back to HDFS.

```
create external table test_cohort_A(test_cohort STRING, event_name STRING, user_id STRING,
                          whisper id STRING, extra information STRING, time generated STRING,
dtype appversion STRING, ts created STRING
row format delimited fields terminated by ',' lines terminated by '\n'
STORED AS TEXTFILE LOCATION 's3n://whisper test/events test cohorts/A';
```

```
INSERT OVERWRITE TABLE test_cohort_A select e.test_cohort, e.event_name, e.user_id, e.whisper_id, e.extra_information, e.time_generated, u.dtype_appversion, u.ts_created from whisper_events e, whisper_users u where e.user_id = u.user_id and e.test_cohort like '%A%';
```

```
""SAME STATEMENT WOULD REPEAT FOR B, C, D, E, F ""
```

We would not have 6 tables created and the files would be at the HDFS location.

I can now easily download these files from S3 using <u>boto</u> for python and run further analysis and enrichments.

<u>SECTION 2: DATA CLEANING AND ENRICHMENT (USING PYTHON AND PANDAS)</u>

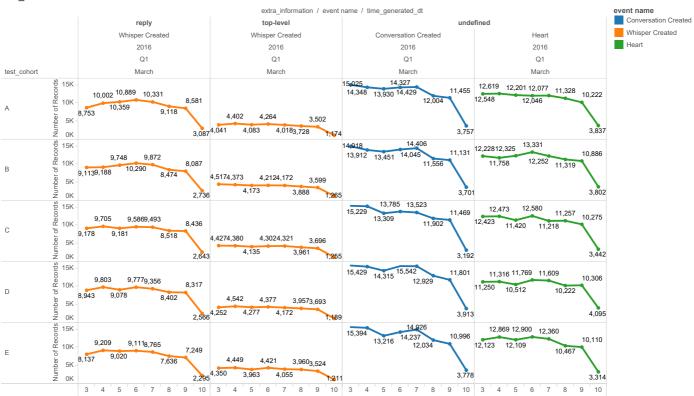
- Convert the events.csv time_generated field to date_time format using pandas. pd.to_datetime(v['time_generated'], unit='ms') (Time in millisecond)
- Convert the users.csv ts_created field to date_time format using pandas. pd.to_datetime(user_df['ts_created'], unit='us')
 (Time in microsecond)

Code Snippet: Function to read csv and enrich it with the conversion to date time using pandas:

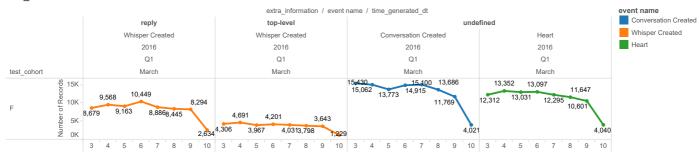
Import pandas as pd

```
def convert_test_cohort_to_dt(file_path_list):
    for file_path in file_path_list:
        df = pd.read_csv(file_path, names=['test_cohort', 'event name', 'user_id', 'whisper_id',
'extra_information', 'time_generated', 'dtype_appversion', 'ts_created'], encoding='utf-8')
        t_generated_dt_df = pd.DataFrame(pd.to_datetime(df['time_generated'], unit='ms'))
        ts_created_dt_df = pd.DataFrame(pd.to_datetime(df['ts_created'], unit='us'))
        t_generated_dt_df = t_generated_dt_df.rename(columns={'time_generated':
'time_generated_dt'))
        ts_created_dt_df = ts_created_dt_df.rename(columns={'ts_created': 'ts_created_dt'})
        df = pd.concat([df, t_generated_dt_df, ts_created_dt_df], axis=1, join_axes=[df.index])
        df[['test_cohort', 'event name', 'user_id', 'whisper_id', 'extra_information', 'time_generated',
'time_generated_dt', 'dtype_appversion', 'ts_created', 'ts_created_dt']].to_csv(file_path+".csv",
index=False)
>
```

SECTION 3: RUNNING ANALYSIS USING TABLEAU



The trend of sum of Number of Records for time_generated_dt Day broken down by extra_information, event name, time_generated_dt Year, time_generated_dt Quarter and time_generated_t Month vs. test_cohort. Color shows details about event name. The marks are labeled by sum of Number of Records. The view is filtered on sum of Number of Records, which ranges from 1,091 to 33,955.



The trend of sum of Number of Records for time_generated_dt Day broken down by extra_information, event name, time_generated_dt Year, time_generated_dt Quarter and time_generated_t Month vs. test_cohort. Color shows details about event name. The marks are labeled by sum of Number of Records. The view is filtered on sum of Number of Records, which ranges from 1,091 to 33,955.

dtype_appversion / event name Number of Records test_cohort N.. android_4.. a.. andro.. andro.. android_4.. android_4.. android_4.. android_4.. android_4.. android_4.. android_4.. android_4.. android_5.. android_5.. android_5.. android_5.. android_5.. android_5.. android_6.. androi 30K 20K 10K 0 0 0 0 0 0 0 000000 0K 30K 20K В 10K 0K 30K Re 20K С 10K 0K 30K 20K D 10K of Records 30K 20K Ε 10K Num 0K -30K of Rec 10K 0K Created versation Created Conversation Created Whisper Created versation Created Whisper Created Whisper Created versation Created versation Created Whisper Whisper Whisper Whisper Whisper Whisper Whisper Whisper

33,955

Sum of Number of Records for each event name broken down by dtype_appversion vs. test_cohort. Color shows sum of Number of Records.

														dty	pe_app	vers	ion / e	ven	t name													
test_cohort		_5.	android_5	5 ar	ndroid	d_5	android_5.	and	droid_5	and	droid_5	.an	droid_5.	.an	droid_5	an	ndroid_	5 a	indroid_5.	.and	droid_5	5an	droid_	_5 a	ndro	id_5.	.and	lroid_5	5 ar	droid_	5 a	ndroid_5
	ති 30K -																															
Λ.	₹ 20K																															
A	한 10K-																															
	30K - 20K -	-0		9-E) (a.,	000	-		9-6		0	OE				2	0.0	٥,	∋⊝ ∈
	8 30K																															
	20K																															
В	ू 10K																															
	d lov		م	عاد	۵۵.					ما						عاد	سمد	ا.						الم	٥۵				عاد	0.0		
	Sp. cor		0 0 0				0 0 0																									
	20K - 30K -																															
С	© 20K-																															
	9 10K																													0.0		
	ਤੋਂ 0K-	-0)E)	··· (-)·		Θ						- 0)	استارياستار	J		- 0)-6)	Θ-	ΘE)(₎	-Θ		9	danahadana)	w/(J⊖€
	5 30K−																															
D	© 20K-																															
	ਰੂ 10K-																												١.			
	∑ 0K	-0	-0-0-0	9€	O-	⊖-		0		0		0)- E)Q(Ð-		-0		9-6) <u>-</u>	Θ-	⊕€)Q	-0	-06	9-10)Q9	2.4	906
	30K -																															
E	₽ 20K																															
	혈 10K																															
	₹ 0K	-0	0-0-0	9-€		⊖-		-0				0		-0	ΘΘ)- C)Q(Ð-						0-	⊖∈		-0		2	209	2.	906
	Number of Records 20K - 10K - 10K - 0K - 10K - 1																															
F	₽ 20K																															
	를 10K																															
	NO P	-0						-0				-0		-0)	۵.		-0				0	OE		-0		2	0.0	0	
		Heart	eated	ated	ated a	Heart	eated eated Heart	ated	eated	ated	eated	ated	eated	pate	eated	pate	ated	Heart	eated eated Heart	ated	eated	ated	ted	Heart	ated	Heart	ated	ated	Deali	ted	Heart	eated
		Ĭ	rersation Created Whisper Created		Whisper Created	Ĭ	rersation Created Whisper Created Heart	Cres	Whisper Created Heart	Cres	Whisper Created	Cres	Whisper Created	Cres	Whisper Created	C.	Whisper Created	Ť	rersation Created Whisper Created Heart	Cre	Whisper Created	Cre	Whisper Created	Ĭ	versation Created Whisper Created	Ĭ	Cre	Whisper Created	Ę į	Whisper Created	Ĭ	Versation Created Whisper Created
			ation	ition	sper		ation	ation	sper	ation	sber	ation	sper	ation	sber	rici	sber		ation	ation	sper	ation	sper		ation	1	ation	sper	ation	sber	-	ation
			Conversation Created Whisper Created	Dearcation Created	Whi		Conversation Created Whisper Created Heart	Conversation Created	W	Conversation Created	W	Conversation Created	W	Conversation Created	N	Conversation Created	N		Conversation Created Whisper Created Heart	Conversation Created	₹	Conversation Created	₹		Conversation Created Whisper Created		Conversation Created	W	Conversation Created	N		Conversation Created Whisper Created
			8	ع ا	3		S	S		S		S		ပ်		Ç	5	- [Š	S		S			Ö		S		اح ا	5	0	3

Sum of Number of Records for each event name broken down by dtype_appversion vs. test_cohort. Color shows sum of Number of Records.

Number of Records

								dty	oe_appversi	on /	event name								
test_cohort		android_5	android_5.	.android_5.	.android_5.	.android_6	ios_3.0.4	ios_3	ios_4.0.1	io	ios_4.1.1	ios_4.2.0	ios_4	ios_4.3.3	ios_4.4.0	ios_4.4.2	io	ios_4	ios_4.
Α	30K Specords Number of Records			000	000														
	10K																		
	∑ 0K	.UOO.	-00-				-0-0-0					G)- -		-0-0-0				
	50 30K				000														
В	₩ 20K			000	0 0														
	ਕੂ 10K			0 0															
	₹ ok	0.0.0	0-0-0			0-0-0	0-0-0					0-0-0)						
	8 30K																		
С	₩ 20K			000	000														
C	o 형 10K			0															
	N OK	0.0.0					-000				-000)						-0
	န် 30K				_														
	£ 20K				000														
D	5 20K			000															
	agun 10K		000									000							
	% 0K										·)				-	0-0	-06
	8 30K				000														
E	£ 20K			000															
	जू 10K			0 0															
	₹ 0K	000	-00				0-0-0	-0-0	0-0-0			-0-0-6)	0-0-0	-	-0			
	წ 30K				000														
F	₽ 20K			000	000														
	호 10K			- 00															
	₽ OK	0.0.0	-0-0-0			0			0-0-0			0-0-0)			0-0-0	,		
		Conversation Created Whisper Created Heart	Conversation Created	Conversation Created Whisper Created Heart	Conversation Created	Conversation Created Whisper Created Heart	Conversation Created Whisper Created	Whisper Created	Conversation Created Whisper Created Heart	Conversation Created Whisper Created Heart	Conversation Created Whisper Created Heart	Heart	Conversation Created Whisper Created	Conversation Created Whisper Created					
		tion (tion (tion (tion (tion (tion (per (tion (tion (tion (tion (tion (
		versa Whis	versa	versa Whis	versa: Whis	versa: Whis	versa	ersa	versa Whis	ersal	versa	versar	Whis	Whis	whis	ersa Whis		ersa' Whis	versa: Whis
		Con	Con	Con	Con	Con	Con	Conv	Con	Conv	Con	Con		Con	Con	Con		Conv	Conv

Sum of Number of Records for each event name broken down by dtype_appversion vs. test_cohort. Color shows sum of Number of Records.

Number of Records

												dty	pe_a	appversi	on / e	event r	name												
test_cohort		.5.2	ios_4.5.3	ios_4.	5.4	ios_4.6.	0 ios_4	4.6.1	ios_4	1.6.2	io	ios_4.6.4	io	s_4.6.5	ios_	4.6.6	ios_5	.1.0	ios_5	.2.0	ios_5	5.4.0	io	ios_5.7.0	ic	os_5.7.	1 i	os_5.8.0	ios_
	Number of Records 30K 10K 0K																												
A	£ 20K																												
	현 10K															_													
	를 0K 용			-00			9					0-0-0)			Ð					-0-0	}⊙-							-0-
	30K																												
}	20K																												
	la 10K			-0-0					0.0			000		0.0		0.0					0.0								
	S over	-		0.0					0-0	,		0.0.0		-0-0		00	0.0				0.0	,							
:	30K																												
	5 20K																												
	Number of Records Number of Re	-0-				O-O-(0-0-0				g@						·····					-		
	န 30K																								T				
	Number of Records 20K 10K 0K																												
	인 10K																												
	₹ 0K		0-0-0	0-0		⊖⊖			-0			0-0-0)		-0-		0-0	⊙-	-00	··· · · · · · · · · · · · · · · · · ·			0				Ð		
	Number of Records 20K 10K 0K																												
	^윤 20K																												
	10K																												
				0-0			Ð		-00	···· · · · · · · · · · · · · · · · · ·		0-0-0)			00		⊖-				···· · · · · · · · · · · · · · · · · ·			-		9-6	900	
	Number of Records 20K 20K 10K 0K																												
:	© 20K																												
	10K																												
	∄ 0K	-O	8 2 E	8 2	t	g ' g ' -	= 2 3	, 'E	8 2)	t	2 2 t			8	g ' ⊑	8 2	⊖-	D D	' ' '	8 8)O-	D	7 2 E)	2 3	8 ' D ' E	8
		Heart	Created Created Heart	Create Create	Heart	Sreate Sreate	Created	Heart	Create	Heart	Heart	Sreated Sreated Heart	Create	Sreated	Create	Sreated	Sreate Sreate	Heart	Sreate Sreate	Heart	Sreate Sreate	Heart	reate	Created Created Heart	Create	Sreated	Look Jes	reated reated Heart	Create
			versation Created Whisper Created Heart	versation Created Whisper Created		versation Created Whisper Created	versation Created	3	Versation Created	i		Versation Created Whisper Created	ation (Whisper Created Heart	ation (Whisper Created	versation Created		versation Created Whisper Created		versation Created	i	Whisper Created	versation Created Whisper Created Heart	ation (Whisper Created	doite	Whisper Created Heart	ation (
			Conversation Created Whisper Created Heart	Conversation Created Whisper Created		Conversation Created Whisper Created	Conversation Created		Conversation Created			Conversation Created Whisper Created Heart	Conversation Created	Whi	Conversation Created	N N	Conversation Created Whisper Created		Conversation Created Whisper Created		Conversation Created Whisper Created		Whi	Conversation Created Whisper Created Heart	Conversation Created	Whi	0,0	Whisper Created Wheart	Conversation Created
			3	S		<u> </u>	S		S			8	ပ်		8		8		S		S			S	0		Č	3	8

Sum of Number of Records for each event name broken down by dtype_appversion vs. test_cohort. Color shows sum of Number of Records.

								dtyp	oe_appversion	on / event n	ame						
est_cohort		s_5.8.1	ios_5.8.2	ios_5.8.3	ios_5.8.4	ios_5.8.5	ios_5.8.6	ios_5.8.7	ios_5.8.8	ios_5.8.9	ios_5.8.10	ios_5.8.11	ios_5.8.12	ios_5.8.13	ios_5.9.0	ios_5.9.1	ios_5.
	Number of Records Number of Re													00			
	20K													0			
	© 10K												000				
	J OK											0.0.0					
	sp		0 0 0	0 0 0	0 0 0	0.0	000	0 0 0	0 0 0	0 0 0	0 0 0	0 0 0		00	0 0 0	0 0 0	
	9 30K													000			
	€ 20K												000				
	용 10K												000				
	≥ 0K	0-0	-0-0-0	0-0-	-0-0-0	0-0-0	0-0-0	-000	0-0-0	0-0-0	-000	-000			00		
	5 30K													000			
	₽ 20K												_				
	호 10K												000				
	₽ OK	0-0		-0-0-0	-0-0-0	-0-0-0		-000	-0-0-0	-0-0-0	0.0.0	0.0.0			-00	0-0-0	-0-6
	န် 30K													0 0			
	[∞] 20K													o o			
	9 10K												000				
	dml ok									0.0.0	0.0.0	000				-0-0-0	
	sp.	0.0	0.0.0	0-0-0	0.0	0.0.0	0-0-0	0.0.0	0.0.0	0.0.0	0.0.0	0.0.0		0 -		0.0.0	
	9 30K													000			
	5 20K												000				
	용 10K												000				
	≥ 0K	-00	-0-0-0	-0-0-0	-0-0-0	0-0-0	0-0-0	-000	0-0-0	-0-0-0	-000	-000			0-0	0-0-0	-06
	9 30K													000			
	₩ 20K																
	ब्रु 10K												000				
	₽ OK	0-0			-0-0-0		0-0-0	-0-0-0	-0-0	-0-0-0	-0-0-0	0.0.0		-		0-0-0	-0-6
		eated	eated	eated	eated	reated eated Heart	reated eated Heart	eated eated Heart	reated	eated eated Heart	eated eated Heart	reated eated Heart	eated eated Heart	eated	reated eated Heart	reated	ated
		Whisper Created Heart	rersation Created Whisper Created Heart	rersation Created Whisper Created Heart	versation Created Whisper Created	versation Created Whisper Created Heart	versation Created Whisper Created Heart	versation Created Whisper Created Heart	rersation Created Whisper Created Heart	rersation Created Whisper Created Heart	rersation Created Whisper Created Heart	rersation Created Whisper Created Heart	/ersation Created Whisper Created Heart	Cres			
		sber	ation	ation	ation	ation	ation	ation	ation	ation	ation	ation	ation	ation	ation	ation	ation
		Whis	Conversation Created Whisper Created Heart	Conversation Created Whisper Created Heart	Conversation Created Whisper Created Heart	Conversation Created Whisper Created Heart	Conversation Created Whisper Created Heart	Conversation Created Whisper Created Heart	Conversation Created Whisper Created Heart	Conversation Created Whisper Created Heart	Conversation Created Whisper Created Heart	Conversation Created Whisper Created Heart	Conversation Created				
			Con	Con	Con	Con	Con	Con	Con	Con	Con	Con	Con	Con	Con	Co	Con

Sum of Number of Records for each event name broken down by dtype_appversion vs. test_cohort. Color shows sum of Number of Records.

Number of Records

			dtyp	e_a	ppv	ersic	n /	eve	nt
test_cohort			.9.2	ios	_5.	9.3	ios	_5.9	9.4
	Number of Records	30K -							
A	of Re	20K -							
	nber	10K -							
	Nur	0K -	-⊖-	0	-0-	·· 🗇 ·	0.	-0-	۰.
	ecords	30K -							
В	of R	20K -							
	nber	10K-							
	Nur	0K -	-⊖-	Θ-	-0-	·· (O)	.0.	-0-	-0-
	ecords	30K -							
С	of R	20K -							
	mber	10K -							
	S Z	0K -	-⊖-	Θ-	-0-	⊖-	.Q.		-0-
	Record	30K -							
D	rofF	20K -							
	nmbe	10K -							
	Z sp	0K -	-⊖-	Θ.	-0-		.Q.	ساريات	-(
	Secore	30K -							
E	rofF	20K -							
	admi	10K -							
	IS N	0K -	-⊖-	Θ-	-0-	·· (O)	-0-		- (-)-
	Secor	30K -							
F	r of F	20K -							
	mbe	10K -							
	ž	0K -	-0-	_	-0-	Θ-	-0-	.0.	
			Heart	Conversation Created	Whisper Created	Heart	Conversation Created	Whisper Created	Heart
				on C	er C		on C	er C	
				ersati	Vhisp		ersati	Vhisp	
				Conve	>		Conve	>	
				-			-		

Number of Records

Sum of Number of Records for each event name broken down by dtype_appversion vs. test_cohort. Color shows sum of Number of Records.

SECTION 4: CONCLUSION

- 1) After running a very detailed analysis on the data and event logs. I can see a very similar pattern in each of the cohort.
- 2) Figure 1: you can see that on the 4th of March the number of whisper <u>replies</u> created increases across all of the cohorts this could be because of the spike in interest because of the change in presentation. I.e. it caused more users to reply to the existing whispers. It is also interesting to note that the number of conversations created also had a downward trend from the 4th till the 5th.
- 3) But each cohort had a similar trend. Which is good that the change affected each cohort group in a similar manner.
- **4) Figure 2:** I wanted to check how the change affected users across devices. Since the logs only contained data from March 3rd till March 11th. We need to check previos logs and see if there is any existing users were affected.
- 5) Focusing on the major groups IOS-5.8.13 noticed a higher whisper creation rate for cohort A, B, and C. While the number of conversations were high across all the cohorts in this group.
- 6) **Figure 3** shown below is optional and gives an hourly trend based on each event_name. It can be a longer and more detailed trend analysis and redundant. But I've always found some people more interested in looking at the subtle differences in each test cohort.
- 7) The most fascinating point is that on the 3rd hour of every day we see a spike in the whispers created, while on the 9th hour there is a minima.

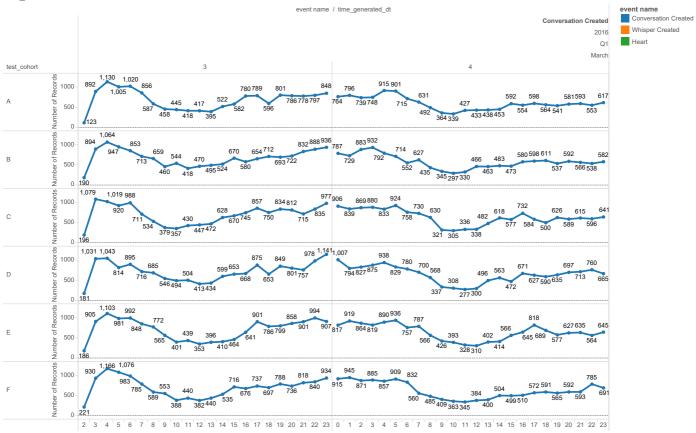
Technology choices:

- 1. A very similar analysis could be produced using R or matlab. I preferred Tableau because I own a licensed version and prefer its strong data visualization properties.
- 2. I enjoy using Python as a scripting language and with data engineering packages like pandas and with the anaconda distro. It just becomes very useable.
- 3. I Used HIVE on AWS to store and partition the data. I always store the data on S3 and create an external table reference from HIVE to the location. In our case we had no partitioned data but if the data was client partitioned or date partitioned its much easier to reference all the data.

THANK	YOU,
OLIVER	LEWIS.

FIGURE 3 BELOW:

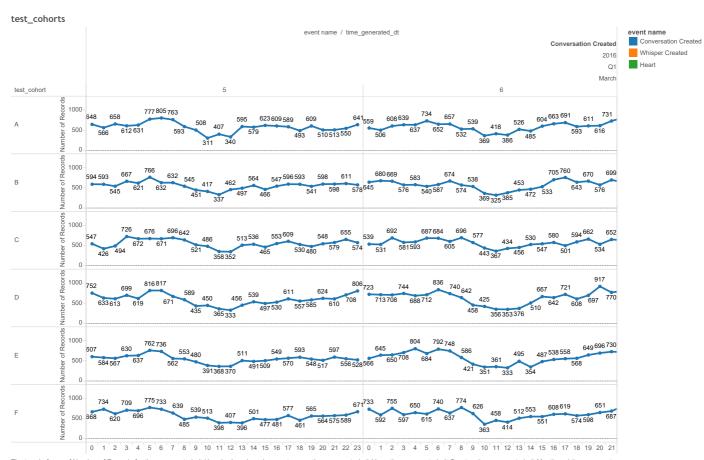




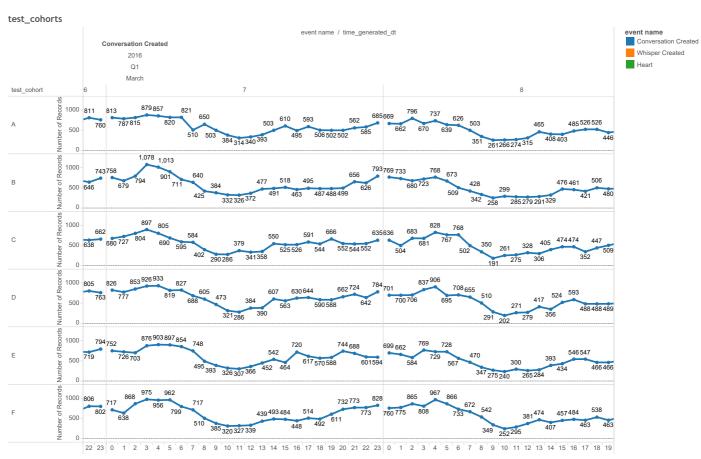
Whisper Created

Heart

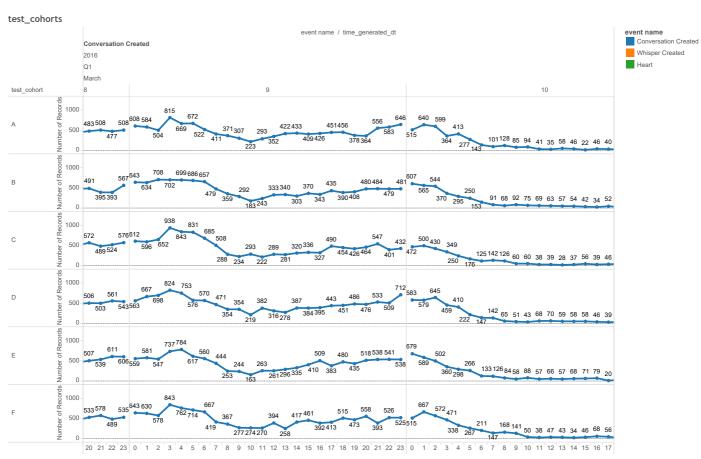
The trend of sum of Number of Records for time_generated_dt Hour broken down by event name, time_generated_dt Year, time_generated_dt Quarter, time_generated_dt Month and time_generated_dt Year, time_generated_dt Year, time_generated_dt Quarter, time_generated_dt Year, ed_dt Day vs. test_cohort. Color shows details about event name. The marks are labeled by sum of Number of Records.



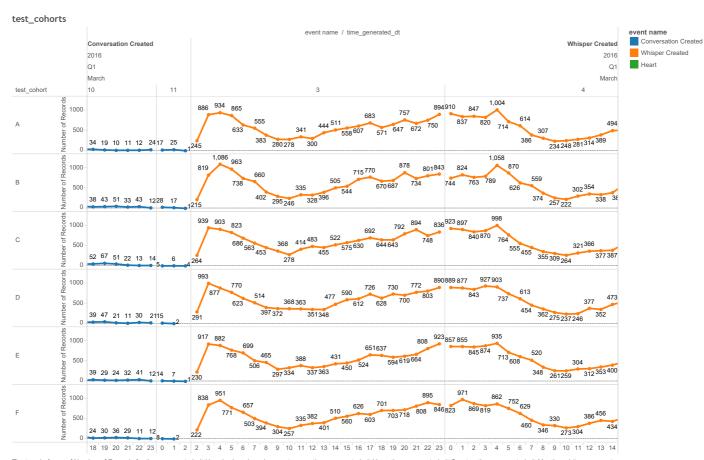
The trend of sum of Number of Records for time_generated_dt Hour broken down by event name, time_generated_dt Year, time_generated_dt Quarter, time_generated_dt Month and time_generated_dt Day vs. test_cohort. Color shows details about event name. The marks are labeled by sum of Number of Records.



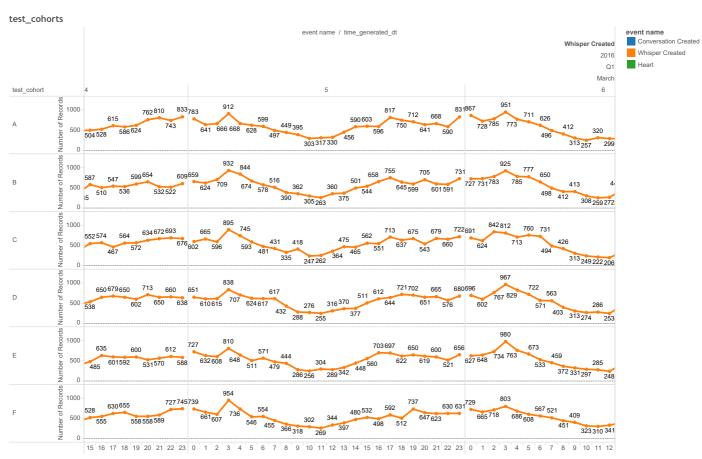
The trend of sum of Number of Records for time_generated_dt Hour broken down by event name, time_generated_dt Year, time_generated_dt Quarter, time_generated_dt Month and time_generated_dt Day vs. test_cohort. Color shows details about event name. The marks are labeled by sum of Number of Records.



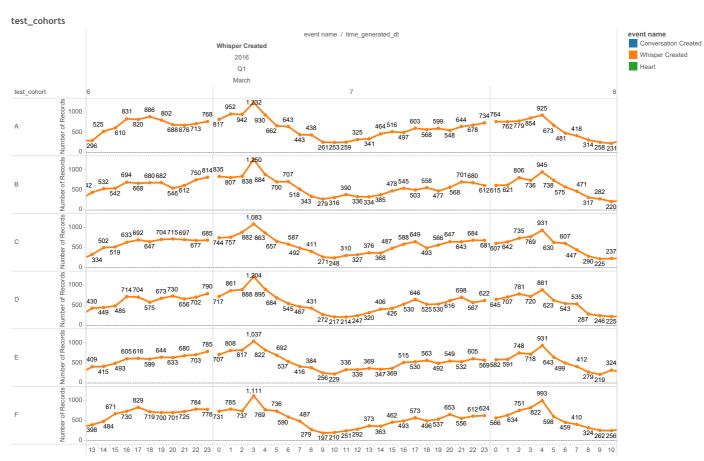
The trend of sum of Number of Records for time_generated_dt Hour broken down by event name, time_generated_dt Year, time_generated_dt Quarter, time_generated_dt Month and time_generated_dt Day vs. test_cohort. Color shows details about event name. The marks are labeled by sum of Number of Records.



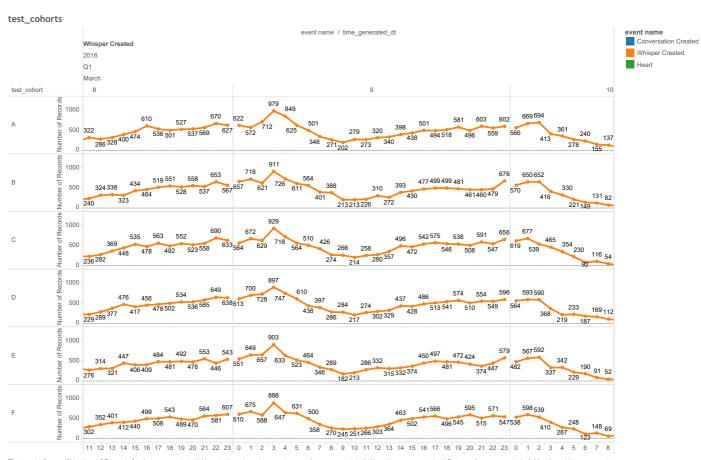
The trend of sum of Number of Records for time_generated_dt Hour broken down by event name, time_generated_dt Year, time_generated_dt Quarter, time_generated_dt Month and time_generated_dt Day vs. test_cohort. Color shows details about event name. The marks are labeled by sum of Number of Records.

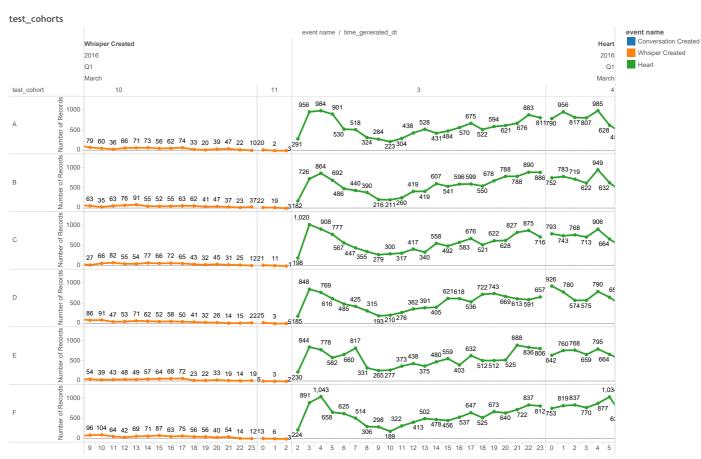


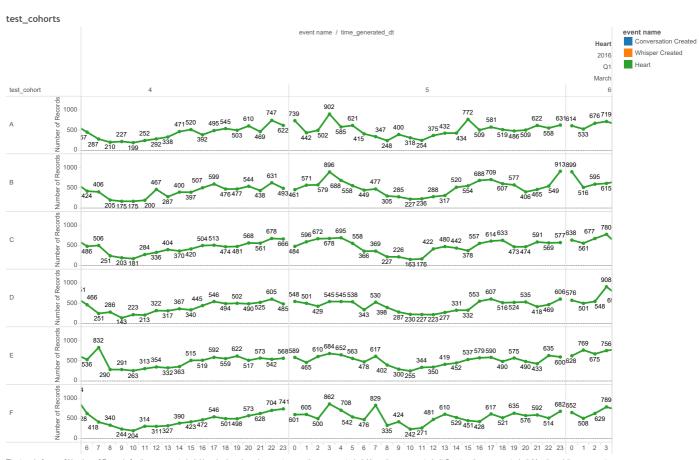
The trend of sum of Number of Records for time_generated_dt Hour broken down by event name, time_generated_dt Year, time_generated_dt Quarter, time_generated_dt Month and time_generated_dt Day vs. test_cohort. Color shows details about event name. The marks are labeled by sum of Number of Records.



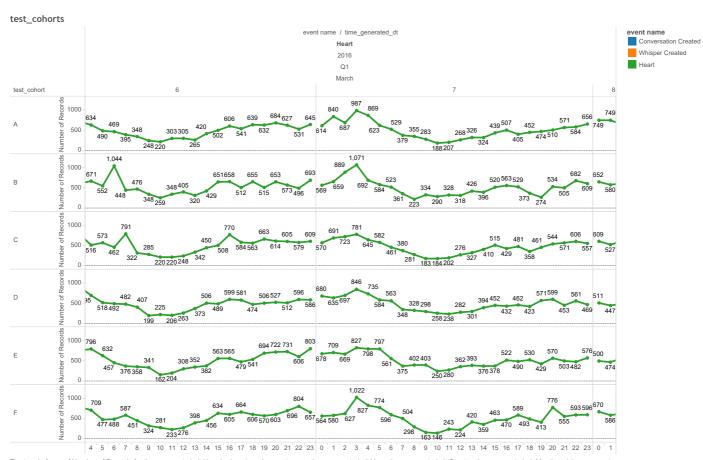
The trend of sum of Number of Records for time_generated_dt Hour broken down by event name, time_generated_dt Year, time_generated_dt Quarter, time_generated_dt Month and time_generated_dt Day vs. test_cohort. Color shows details about event name. The marks are labeled by sum of Number of Records.

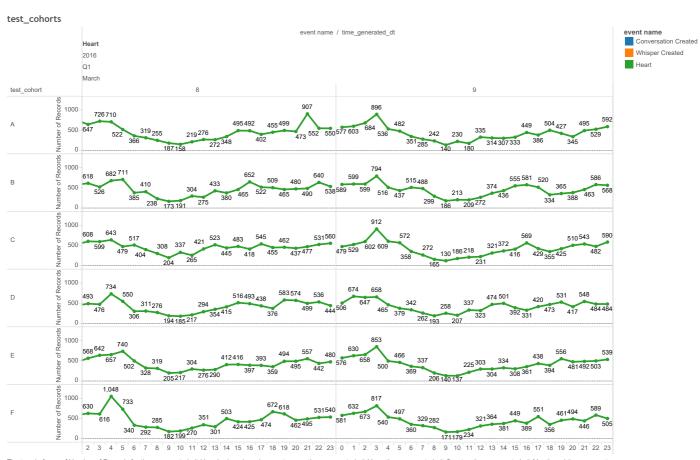






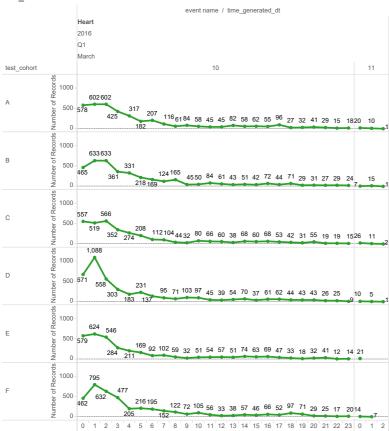
The trend of sum of Number of Records for time_generated_dt Hour broken down by event name, time_generated_dt Year, time_generated_dt Quarter, time_generated_dt Month and time_generated_dt Day vs. test_cohort. Color shows details about event name. The marks are labeled by sum of Number of Records.





The trend of sum of Number of Records for time_generated_dt Hour broken down by event name, time_generated_dt Year, time_generated_dt Quarter, time_generated_dt Month and time_generated_dt Day vs. test_cohort. Color shows details about event name. The marks are labeled by sum of Number of Records.





event name Conversation Created Whisper Created

Heart