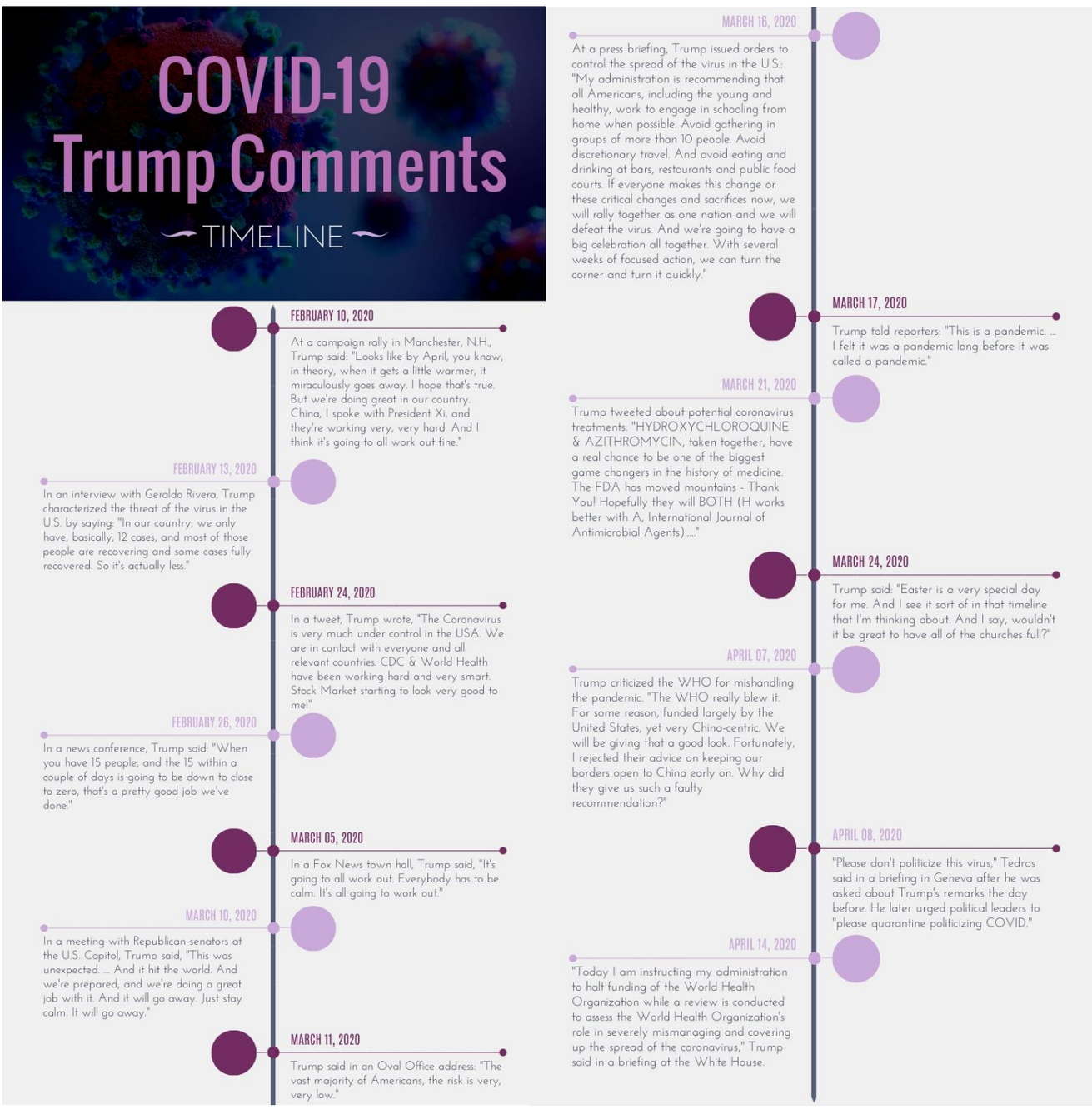
Extract Transform Load

explained

במשימת הלמידה שלנו, ננסה לחזות האם טוויטים מדברים על נשיא ארה"ב **Donald Trump** בתאריכים שונים בהם הוא הגיב באופן ציבורי לגבי נושאים הנוגעים למחלת הקורונה.
לשם כך, נניח כי טוויטים שנושאם אכן קשור אל הנשיא **Trump**, יכילו תיוג שמכיל את המילה **Trump** בתוכו.

למציאת אותם תאריכים, נעזרנו במידע מהאינטרנט המציג את הצהרותיו ותגובותיו אל הנושא בתאריכים שונים.
בהתאם לכך, המידע שנחלץ עבור משימת הלמידה שלנו, יתבסס על ה **topics** הנוגעים לתאריכים אלו המוצגים בתרשים הבא:



מקור:

A Timeline Of Coronavirus Comments From President Trump And WHO, ***By Tamara Keith and Malaka Gharib, NPR.***

Data store description and design:

לביצוע המשימה בחרנו להשתמש ב **Elasticsearch** לאחסון המידע וניתוחו.
בחרנו 14 תאריכים שונים (המצוינים לעיל) שישמשו אותנו לביצוע המשימה וכתבנו אותם אל **Elasticsearch** בזוגות. כלומר, כל 2 תאריכים נכתבו יחדיו אל אותו אינדקס כך שסך הכל עבדנו עם 7 אינדקסים. ביזור המידע בין האינדקסים בצורה זו אפשר לנו לכתוב את המידע בצורה מקבילית (ומכך מהירה יותר) אל **Elasticsearch**. בעזרת הכלים של **Kibana** יצרנו **Index pattern** מתאים הכולל את המידע מכל 14 התאריכים.
לכתיבת הנתונים בצורה מקבילית כזו, הרצנו קוד זהה ב7 מחברות שונות (עם 7 מפתחות שונים מטוויטר) ובכל מחברת כתבנו לאינדקס שונה. המחברת המצורפת מציגה כתיבה יחידה אל אינדקס ספציפי (ההבדל אל שאר המחברות מתבטא בהבדל בין התאריכים והמפתחות מטוויטר בלבד).

להלן המידע שקלטנו אל **Elasticsearch** באינדקסים שונים:

Name	Health	Status	Primaries	Replicas	Docs count	Storage size
trump_covid-19_tweets_24-03-2020_07-04-2020	green	open	1	0	1997572	121.6mb
trump_covid-19_tweets_24-02-2020_26-02-2020	green	open	1	0	1587796	90.2mb
trump_covid-19_tweets_17-03-2020_21-03-2020	green	open	1	0	2484995	138.4mb
trump_covid-19_tweets_11-03-2020_16-03-2020_2nd_try	green	open	1	0	3519538	188.4mb
trump_covid-19_tweets_10-02-2020_13-02-2020	green	open	1	0	417598	26.1mb
trump_covid-19_tweets_08-04-2020_14-04-2020_2	green	open	1	0	2592177	161.7mb
trump_covid-19_tweets_05-03-2020_10-03-2020	green	open	1	0	1988252	106.3mb

שם האינדקס המכיל את המידע הרלוונטיים עבור התאריכים **X,Y** נקרא **trump_covid-19_tweets_X_Y**. בחלק מהמקרים נאלצנו להריץ את כתיבת הנתונים בשנית בשל בעיות תקשורת ובהתאם יצרנו אינדקס חדש עם שם שונה במעט (על מנת לא למחוק את הנתונים באינדקס המקורי).

ב-**Settings** של כל אינדקס בחרנו להשתמש ב-**Shard** יחיד לכל אינדקס מכיוון שבמקורות שקראנו של **Elasticsearch** ההמלצה היא להוסיף **Shard** על כל **20GB** של נתונים.
היות וכתבנו כל זוג תאריכים לאינדקס נפרד נפח הנתונים שנכתבו לא הגיע לכמות המומלצת ולכן השארנו **Shard** אחד לכל אינדקס.

קישור למחברת ה ETL בDATABRICS