BUILDING A RELATIONAL ONLINE ANALYTICAL PROCESSING SYSTEM

YouTube And TED Talks Datasets

Written by
Nimisha Budhwani, Faridun Rakhmonov, Herve Tchouamani
Supervised by Professor Holowczak

Abstract

This paper aims to integrate two system of records into a Data Warehouse following Kimball methodology. We have defined six Key Performance Indicators to support data sources that are represented by YouTube videos and TED talks from 2006 to 2018. SQL Server has been chosen as a target database to support data marts, and it is hosted by Microsoft Azure. Before running the ETL, some dimensions have been cleaned using Tableau Prep. The ETL process has been implemented on eleven of the thirty-four subsystems as defined by Kimball. After loading the data into the Data Warehouse, Tableau has been connected to Microsoft SQL Server, and we came up with meaningful insights from the data using some visualization based on comparative analysis, trending analysis, or correlation analysis.

Key Words: Data Integration, ETL, Cloud Computing, Relational Databases, Business Intelligence, Key Performance Indicators

Introduction

YouTube is a world-famous video sharing website, that accounts for 11.4% web traffic alone, making it the third highest website in terms of web traffic. On the other side, TED Talks began in 1984 as a conference, and today covers almost all topics from science to business to global issues in more than 100 languages. In 2012, TED Talks had been watched over 1 billion times worldwide. With such a huge number of audiences which is continually growing, it is essential to find out what is bringing this audience to the platform, what kind of topics attract the maximum discussion in the form of comments? YouTube, being a competitor of TED Talk in video sharing space it is vital to understand if the business model for TED Talk is sustainable in terms of its longevity. By being able to analyze how and why YouTube has been successful it will be beneficial in understanding what kind of content is attracting such a huge number of audiences for YouTube.

To provide an accurate answer to these concerns, we are going to integrate different source systems to an Enterprise Data Warehouse for analytic purposes. This Data Warehouse will highly depend on the business requirements set up upstream. That is the reason why the first section of this project is to define some Key Performance Indicators (KPI) which dictate the granularity of the data and therefore the type of data needed to achieve the goal. The next section is to set up the technical architecture by deciding on the type of the target database to support the Data Warehouse. In addition, we also need to decide if the Data Warehouse will be locally hosted or cloud based. Then, the third section offers details regarding the data integration in which we present the physical design and the ETL process. Finally, the last section, which is the business intelligence provides knowledges, meaningful insights, patterns we can extract from the data.

I. Business Requirement: Key Performance Indicators (KPIs)

To get a sense of how our business performing, we identified six KPIs that measure values reporting progress against our desired results. Each of the KPIs represents requirements that inspire some kind of actions in making progress. Once we collect data from each of the KPIs, it would help the business in making decisions to posting/storing the types of videos that are performing good or bad in their websites.

KPI #1: *Top 10 Videos by Views*

- What kind of videos are performing the best in terms of views?

KPI #2: *Video Engagement*

1. TED Talk:

    Number of Views

    Number of Comments

2. YouTube:

    Number of Views

    Number of Likes

    Number of Dislikes

    Number of Comments

KPI #3: *Average Daily Views*

- What days are popular/unpopular when watching YouTube Videos or TED Talk videos?
- When is it the best day to upload a video on YouTube/TED Talk?

KPI #4: *Percentage of likes based on total views*

- How many people are reacting to each YouTube video?

KPI #5: *Top 3 Popular Topics viewed on* TED Talk/YouTube

- What topics are attracting the most audience?

KPI #6: *the holiday day that has the most views.*

II. Technical Architecture Design and Product Selection.

In this section, we are going to focus on the data source, the information access and data integration, and the Datawarehouse. We will have to decide to keep a technology that we already know, such as a relational OLAP or to go for something new like a multidimensional OLAP or a hybrid system.

Data Source

We downloaded two datasets on Kaggle. The first one is a YouTube dataset that possesses a CSV file and a JSON file. The CSV file contains information about videos published in the United States. It has 45949 rows and 16 fields among which we have a Category ID field that stores the ID of the topic of each video. This field is fully described in a JSON where each ID is associated with its Category Name. For each video in the CSV file, we have measurements regarding the number of comments, views, dislikes, and likes recorded at the trending date. The second dataset is a TED Talks dataset where data is observed in the United States and stored in a CSV file with 2550 rows. There are 17 attributes that describe each Talk and we can assess the performance of each talk by the number of views and the number of comments.

Information Access and Data Integration

There are various ways to integrate and access the YouTube and TED Talk dataset. In order to integrate these two datasets, we had to conform to a fixed schema by using dimensional modeling technique. In order to integrate the source systems, we used middleware APIs specifically the JDBC drive when connecting to the target database in Pentaho Data Integration.

Data Warehouse/Data Marts

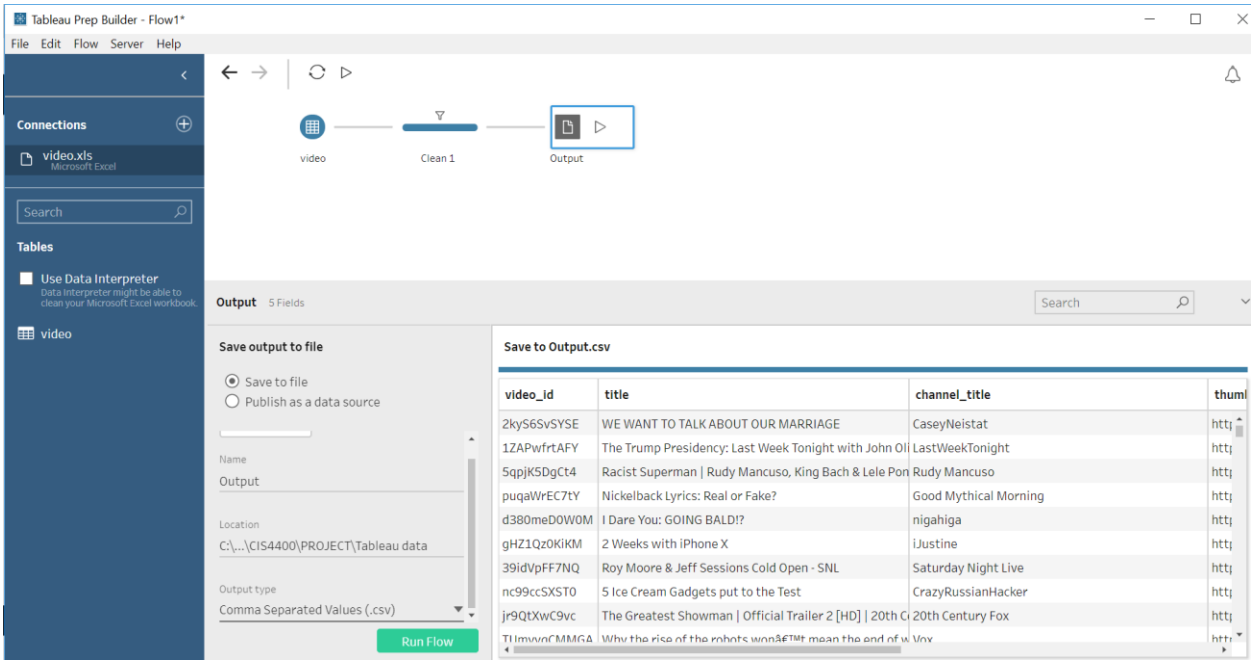
As a target database we decided to use Microsoft SQL Server database which will be hosted in one of the best cloud services providers, Microsoft AZURE. This database can be classified as a Relational Online Analytical Processing (ROLAP). Since we had a familiarity working with relational databases, it gave us an advantage. Another beneficial aspect about working with relational database is the fact that it uses SQL as a data definition language. The database will use a fixed schema. Since we are performing Extract, Transform, and Load (ETL), we have to structure the data before loading it into the target database. The reason for building this system is for analytical purposes.

III. Data Integration

This section will allow us to present the Data Profiling step which is very important for the accuracy of the analysis. In the dimensional modeling part, we are just going to describe dimensions and fact tables that we came up with. Then, the physical design will cover the next paragraph followed by the ETL process.

Data Profiling

The video dimension had some inappropriate values in some of its fields. The video_id and description fields contained “null” values. This dimension has been loaded into Tableau Prep to filter out all null values. After applying the filter, the output file has been downloaded in a xls file and then transformed into a csv file in order to be sent through the ETL pipeline. The transformation is shown in the figure below.

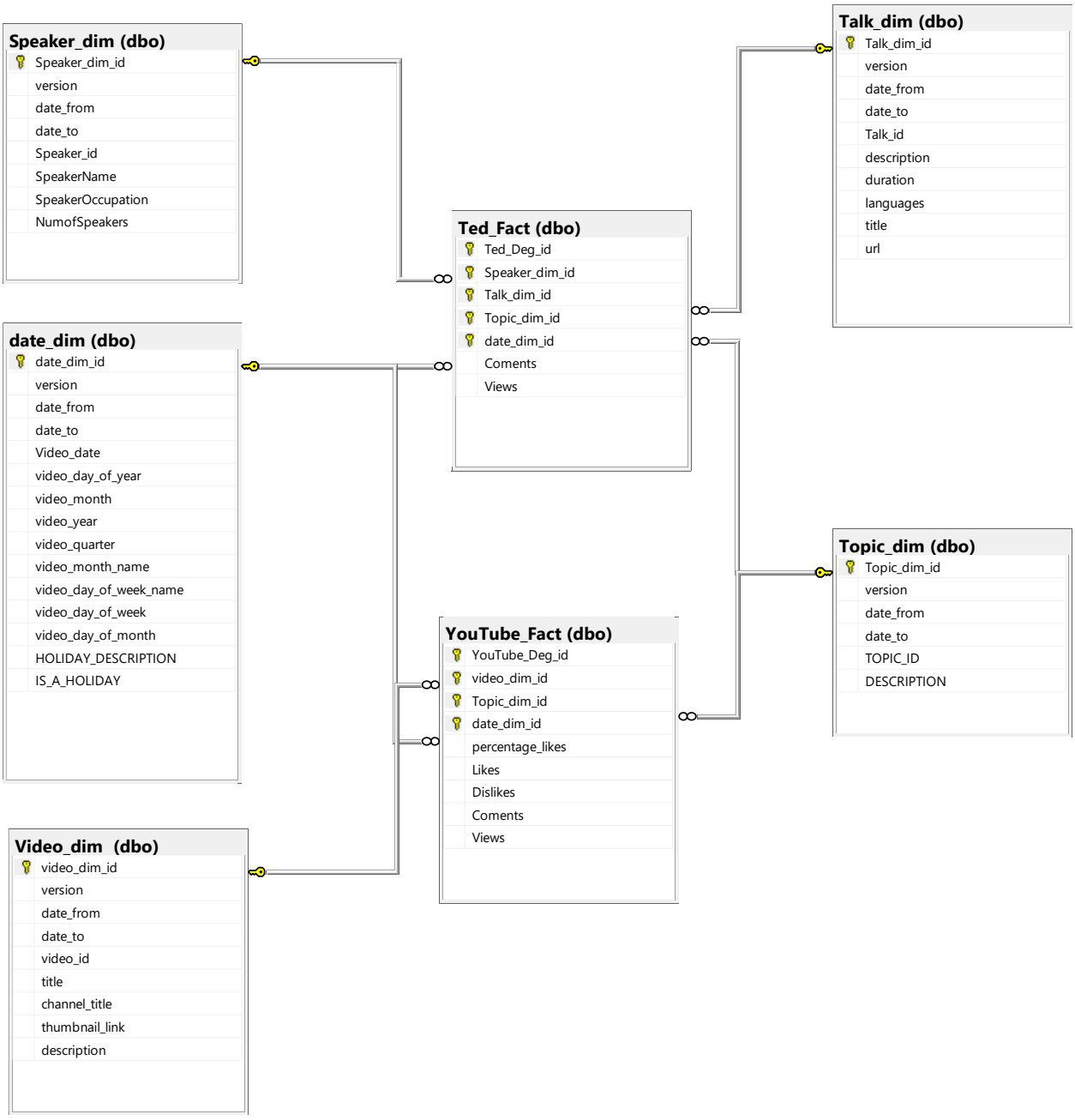


List of Dimensions and Fact Tables Identified

For this project, we expect to build a data warehouse with **five** dimensions and **two** fact tables. The dimension supplies the context under which a fact table is true, so we will have a **Talks** Dimension, **Topic** Dimension, **Speaker** Dimension, and **Date** Dimension. These dimensions will make the **TED** Fact Table. The measurements that are part of this fact table are the number of views and the number of comments. For the YouTube dataset, we established the **Video** Dimension, **Topic** Dimension, and **Date** Dimension. This will make up the **YouTube** Fact Table. The measurements that will be part of this fact table are the number of views, number of likes, number of dislikes and the number of comments. The **Topic** and **Date** dimension will be used as conformed dimensions to drill across both data Marts.

Physical Design of the ROLAP.

SQL Server Management Studio is the client used to connect to the Database hosted in Microsoft AZURE. It is also the tool used to draw the diagram below which represents the physical design in a constellation schema. This schema has two fact tables that share the Topic and Date dimension. The attributes of the Date dimension form a hierarchy. Each dimension Talks, Speaker, Video, and Topic has two ID's: the first ID is the surrogate key that uniquely identifies each row and the second ID represents the natural key from the OLTP systems. The primary key of the TED and YouTube fact tables is the concatenation of the foreign keys from each dimension. In addition, TED_Deg_id and YouTube_Deg_id are degenerate keys that need to be combined with foreign keys in order to allow every single row to be uniquely identified. To create the relationship between each dimension and the fact table, the referential integrity has been enforced. Each surrogate key in the fact table must be a primary key in another table.

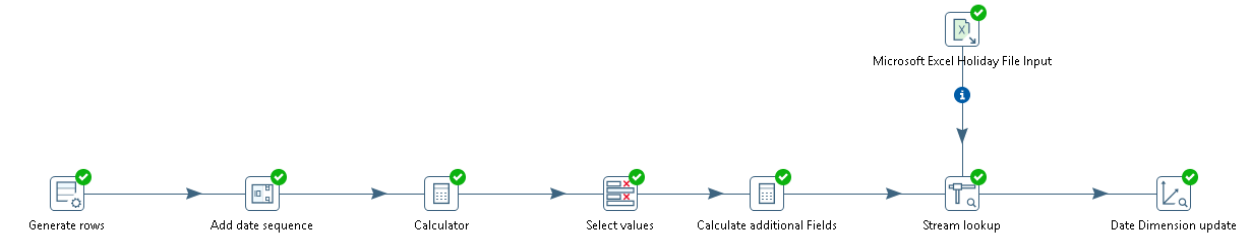


Extraction Transformation and Loading (ETL)

1. Date Dimension

The Date Dimension is a conformed dimension in which columns include video_date, video_day_of_year, video_month, video_quarter, video_month_name,

video_day_of_week_name, video_day_of_week, video_day_of_month, holiday_description, and is_a_holiday. When 'Generate Row' we specified the limit to 4,015 since we have about 11 years of data. We assigned 'date_dim_id' as the surrogate key for date dimension. Through 'Calculate Additional Fields' we are able to embellish based on week, month, quarter, year and holiday. Since we don't need to change the date dimension after the initial load, we have implemented a type zero slowly changing dimension. The Holiday file input contains all the holidays from years 2006-2018. By adding a dimensional hierarchy allows us to drill up and down in order to gain a better analysis about the data. Since this dimension is shared with the Youtube and TED Talk fact table it also helps us drill across.



Execution Results

Execution History | Logging | Step Metrics | Performance Graph | Metrics | Preview data

#	Stepname	Copynr	Read	Written	Input	Output	Updated	Rejected	Errors	Active	Time	Speed (r/s)	input/output
1	Generate rows	0	0	4760	0	0	0	0	0	Finished	0.0s	528,889	-
2	Add date sequence	0	4760	4760	0	0	0	0	0	Finished	0.0s	238,000	-
3	Calculator	0	4760	4760	0	0	0	0	0	Finished	0.0s	105,778	-
4	Select values	0	4760	4760	0	0	0	0	0	Finished	0.0s	97,143	-
5	Microsoft Excel Holiday File Input	0	0	0	0	0	0	0	0	Finished	0.0s	0	-
6	Calculate additional Fields	0	4760	4760	0	0	0	0	0	Finished	0.2s	27,514	-
7	Stream lookup	0	4760	4760	0	0	0	0	0	Finished	0.2s	26,742	-
8	Date Dimension update	0	4760	4760	4760	4760	0	0	0	Finished	3mn 59s	20	-

Examine preview data

Rows of step: date_dim (100 rows)

#	date_dim_id	version	date_from	date_to	Video_date	video_day_of_year	video_month	video_year	video_quarter	video_month_name	video_day_of_week_name	video_day_of_week	video_day_of_month	HOLIDAY_DESCRIPTION	IS_A_HOLIDAY
43	42	1	1900/01/01 00...	2200/01/01 ...	2006/04/08 ...	98	4	2006	2	April	Saturday	7	8	<null>	0
44	43	1	1900/01/01 00...	2200/01/01 ...	2006/04/09 ...	99	4	2006	2	April	Sunday	1	9	<null>	0
45	44	1	1900/01/01 00...	2200/01/01 ...	2006/04/10 ...	100	4	2006	2	April	Monday	2	10	<null>	0
46	45	1	1900/01/01 00...	2200/01/01 ...	2006/04/11 ...	101	4	2006	2	April	Tuesday	3	11	<null>	0
47	46	1	1900/01/01 00...	2200/01/01 ...	2006/04/12 ...	102	4	2006	2	April	Wednesday	4	12	<null>	0
48	47	1	1900/01/01 00...	2200/01/01 ...	2006/04/13 ...	103	4	2006	2	April	Thursday	5	13	<null>	0
49	48	1	1900/01/01 00...	2200/01/01 ...	2006/04/14 ...	104	4	2006	2	April	Friday	6	14	Good Friday	1
50	49	1	1900/01/01 00...	2200/01/01 ...	2006/04/15 ...	105	4	2006	2	April	Saturday	7	15	<null>	0
51	50	1	1900/01/01 00...	2200/01/01 ...	2006/04/16 ...	106	4	2006	2	April	Sunday	1	16	<null>	0
52	51	1	1900/01/01 00...	2200/01/01 ...	2006/04/17 ...	107	4	2006	2	April	Monday	2	17	<null>	0
53	52	1	1900/01/01 00...	2200/01/01 ...	2006/04/18 ...	108	4	2006	2	April	Tuesday	3	18	<null>	0
54	53	1	1900/01/01 00...	2200/01/01 ...	2006/04/19 ...	109	4	2006	2	April	Wednesday	4	19	<null>	0
55	54	1	1900/01/01 00...	2200/01/01 ...	2006/04/20 ...	110	4	2006	2	April	Thursday	5	20	<null>	0
56	55	1	1900/01/01 00...	2200/01/01 ...	2006/04/21 ...	111	4	2006	2	April	Friday	6	21	<null>	0
57	56	1	1900/01/01 00...	2200/01/01 ...	2006/04/22 ...	112	4	2006	2	April	Saturday	7	22	<null>	0
58	57	1	1900/01/01 00...	2200/01/01 ...	2006/04/23 ...	113	4	2006	2	April	Sunday	1	23	<null>	0
59	58	1	1900/01/01 00...	2200/01/01 ...	2006/04/24 ...	114	4	2006	2	April	Monday	2	24	<null>	0
60	59	1	1900/01/01 00...	2200/01/01 ...	2006/04/25 ...	115	4	2006	2	April	Tuesday	3	25	<null>	0

2. Speaker Dimension

Dimensions supply the context for the facts. Before inputting the CSV file, we had to create individual excel file with all the attributes that exists in the speaker dimensions. To form any dimensions, we take tables from the Online Transactional Processing (OLTP) system and perform the following steps: De-normalize data, embellish, and add a surrogate key. The surrogate key for the speaker dimension is named 'speaker_dim_id'. In Pentaho this is known as the 'technical key'. The native key that we bring from the source here is the 'Speaker_id'. This is essential in connecting the source to the target. We were able to embellish it by adding occupation of the speaker and the number of speakers. To keep track of the historical data if a new speaker is added or if a speaker changes his name, the type 2 slowly changing dimension has been implemented.

Execution Results

Execution History | Logging | Step Metrics | Performance Graph | Metrics | Preview data

#	Stepname	Copynr	Read	Written	Input	Output	Updated	Rejected	Errors	Active	Time	Speed (r/s)	input/output
1	Speaker CSV file input	0	0	2374	2375	0	0	0	0	Finished	0.0s	53,977	-
2	Speaker Dimension lookup/update	0	2374	2374	2374	2374	0	0	0	Finished	2mn 0s	20	-

Examine preview data

Rows of step: Speaker_dim (50 rows)

#	Speaker_dim_id	version	date_from	date_to	Speaker_id	SpeakerName	SpeakerOccupation	NumofSpeakers
1	0	1	<null>	<null>	<null>	<null>	<null>	<null>
2	1	1	2006/01/01 00:00:00.000000000	2200/01/01 00:00:00.000000000	SP100	Ken Robinson	Author/educator	<null>
3	2	1	2006/01/01 00:00:00.000000000	2200/01/01 00:00:00.000000000	SP101	Al Gore	Climate advocate	<null>
4	3	1	2006/01/01 00:00:00.000000000	2200/01/01 00:00:00.000000000	SP102	David Pogue	Technology columnist	<null>
5	4	1	2006/01/01 00:00:00.000000000	2200/01/01 00:00:00.000000000	SP103	Majora Carter	Activist for environmental justice	<null>
6	5	1	2006/01/01 00:00:00.000000000	2200/01/01 00:00:00.000000000	SP104	Hans Rosling	Global health expert; data visionary	<null>
7	6	1	2006/01/01 00:00:00.000000000	2200/01/01 00:00:00.000000000	SP105	Tony Robbins	Life coach; expert in leadership psychology	<null>
8	7	1	2006/01/01 00:00:00.000000000	2200/01/01 00:00:00.000000000	SP106	Julia Sweeney	Actor, comedian, playwright	<null>
9	8	1	2006/01/01 00:00:00.000000000	2200/01/01 00:00:00.000000000	SP107	Joshua Prince-Ramus	Architect	<null>
10	9	1	2006/01/01 00:00:00.000000000	2200/01/01 00:00:00.000000000	SP108	Dan Dennett	Philosopher, cognitive scientist	<null>
11	10	1	2006/01/01 00:00:00.000000000	2200/01/01 00:00:00.000000000	SP109	Rick Warren	Pastor, author	<null>
12	11	1	2006/01/01 00:00:00.000000000	2200/01/01 00:00:00.000000000	SP110	Cameron Sinclair	Co-founder, Architecture for Humanity	<null>
13	12	1	2006/01/01 00:00:00.000000000	2200/01/01 00:00:00.000000000	SP111	Jehane Noujaim	Filmmaker	<null>
14	13	1	2006/01/01 00:00:00.000000000	2200/01/01 00:00:00.000000000	SP112	Larry Brilliant	Epidemiologist, philanthropist	<null>
15	14	1	2006/01/01 00:00:00.000000000	2200/01/01 00:00:00.000000000	SP113	Jeff Han	Human-computer interface designer	<null>
16	15	1	2006/01/01 00:00:00.000000000	2200/01/01 00:00:00.000000000	SP114	Nicholas Negroponte	Tech visionary	<null>
17	16	1	2006/01/01 00:00:00.000000000	2200/01/01 00:00:00.000000000	SP115	Sirena Huang	Violinist	<null>
18	17	1	2006/01/01 00:00:00.000000000	2200/01/01 00:00:00.000000000	SP116	Amy Smith	inventor, engineer	<null>
19	18	1	2006/01/01 00:00:00.000000000	2200/01/01 00:00:00.000000000	SP117	Richard Baraniuk	Education visionary	<null>
20	19	1	2006/01/01 00:00:00.000000000	2200/01/01 00:00:00.000000000	SP118	Mena Trott	Blogger; cofounder, Six Apart	<null>

3. Talk Dimension

The surrogate key in the Talk Dimension is represented by ‘Talk_dim_id’ and the native key is represented by ‘Talk_id’. The talk dimension consists of description of the video, the number of languages that the video was translated in, the title of the video, and url. Any update in this dimension will keep the historical data since the type 2 slowly changing dimensional has been set up.

Execution Results

Execution History | Logging | Step Metrics | Performance Graph | Metrics | Preview data

#	Stepname	Copynr	Read	Written	Input	Output	Updated	Rejected	Errors	Active	Time	Speed (r/s)	input/output
1	Talk CSV file input	0	0	2374	2375	0	0	0	0	Finished	0.1s	39,583	-
2	Talk Dimension update	0	2374	2374	2374	2374	0	0	0	Finished	2mn 7s	19	-

Talk_dim_id	version	date_from	date_to	Talk_id	description	duration	languages	title	url
0	1	<null>	<null>	<null>	<null>	<null>	<null>	<null>	<null>
1	1	2006/01/01 ...	2200/01/01 00...	TA100	Sir Ken Robinson ...	1164	60	Do schools kill creativity?	https://www.ted.com/talks/ken_robinson_says_schools_kill_creat...
2	1	2006/01/01 ...	2200/01/01 00...	TA101	With the same hu...	977	43	Averting the climate crisis	https://www.ted.com/talks/al_gore_on_averting_climate_crisis
3	1	2006/01/01 ...	2200/01/01 00...	TA102	New York Times c...	1286	26	Simplicity sells	https://www.ted.com/talks/david_pogue_says_simplicity_sells
4	1	2006/01/01 ...	2200/01/01 00...	TA103	In an emotionally...	1116	35	Greening the ghetto	https://www.ted.com/talks/majora_carter_s_tale_of_urban_renewal
5	1	2006/01/01 ...	2200/01/01 00...	TA104	You've never seen...	1190	48	The best stats you've ever seen	https://www.ted.com/talks/hans_rosling_shows_the_best_stats_y...
6	1	2006/01/01 ...	2200/01/01 00...	TA105	Tony Robbins dis...	1305	36	Why we do what we do	https://www.ted.com/talks/tony_robbins_asks_why_we_do_what...
7	1	2006/01/01 ...	2200/01/01 00...	TA106	When two young...	992	31	Letting go of God	https://www.ted.com/talks/julia_sweeney_on_letting_go_of_god
8	1	2006/01/01 ...	2200/01/01 00...	TA107	Architect Joshua ...	1198	19	Behind the design of Seattle's library	https://www.ted.com/talks/joshua_prince_ramus_on_seattle_s_li...
9	1	2006/01/01 ...	2200/01/01 00...	TA108	Philosopher Dan ...	1485	32	Let's teach religion -- all religion -- in s...	https://www.ted.com/talks/dan_dennett_s_response_to_rick_war...
10	1	2006/01/01 ...	2200/01/01 00...	TA109	Pastor Rick Warre...	1262	31	A life of purpose	https://www.ted.com/talks/rick_warren_on_a_life_of_purpose
11	1	2006/01/01 ...	2200/01/01 00...	TA110	Accepting his 200...	1414	27	My wish: A call for open-source archite...	https://www.ted.com/talks/cameron_sinclair_on_open_source_ar...
12	1	2006/01/01 ...	2200/01/01 00...	TA111	Jehane Noujaim u...	1538	20	My wish: A global day of film	https://www.ted.com/talks/jehane_noujaim_inspires_a_global_da...
13	1	2006/01/01 ...	2200/01/01 00...	TA112	Accepting the 20...	1550	24	My wish: Help me stop pandemics	https://www.ted.com/talks/larry_brilliant_wants_to_stop_pande...
14	1	2006/01/01 ...	2200/01/01 00...	TA113	Jeff Han shows of...	527	27	The radical promise of the multi-touch ...	https://www.ted.com/talks/jeff_han_demos_his_breakthrough_t...
15	1	2006/01/01 ...	2200/01/01 00...	TA114	Nicholas Negrop...	1057	25	One Laptop per Child	https://www.ted.com/talks/nicholas_negroponte_on_one_laptop...
16	1	2006/01/01 ...	2200/01/01 00...	TA115	Violinist Sirena H...	1481	31	An 11-year-old's magical violin	https://www.ted.com/talks/sirena_huang_dazzles_on_violin

4. Topic Dimension

Topic Dimension is a conformed dimension because it helps in connecting both Youtube dataset and TED Talk dataset. This is essential during analysis because it helps drill across. This dimension corresponds to all the topic videos that are posted on these two platforms. The surrogate key here is 'Topic_dim_id' and the native key that is bought from the OLTP is the 'Topic_id'. Including the topic dimension helps assign each video to a category and any update will be capture according to the type 2 slowly changing dimension.

Execution Results

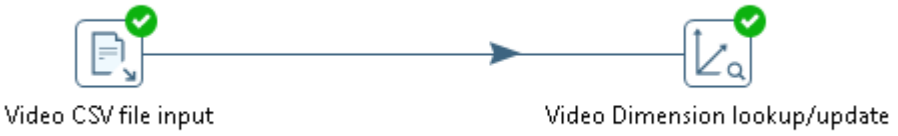
Execution History | Logging | Step Metrics | Performance Graph | Metrics | Preview data

#	Stepname	Copynr	Read	Written	Input	Output	Updated	Rejected	Errors	Active	Time	Speed (r/s)	input/output
1	Topic CSV file input	0	0	31	32	0	0	0	0	Finished	0.0s	8,000	-
2	Topic Dimension lookup/update	0	31	31	31	31	0	0	0	Finished	1.6s	20	-

#	Topic_dim_id	version	date_from	date_to	TOPIC_ID	DESCRIPTION
1	0	1	<null>	<null>	<null>	<null>
2	1	1	2006/01/01 00:00:00.000000000	2200/01/01 00:00:00.000000000	1	animation/film
3	2	1	2006/01/01 00:00:00.000000000	2200/01/01 00:00:00.000000000	2	autos/cars/vehicles
4	3	1	2006/01/01 00:00:00.000000000	2200/01/01 00:00:00.000000000	10	dance/music
5	4	1	2006/01/01 00:00:00.000000000	2200/01/01 00:00:00.000000000	15	animals/pets
6	5	1	2006/01/01 00:00:00.000000000	2200/01/01 00:00:00.000000000	17	sports
7	6	1	2006/01/01 00:00:00.000000000	2200/01/01 00:00:00.000000000	18	short movies
8	7	1	2006/01/01 00:00:00.000000000	2200/01/01 00:00:00.000000000	19	events/exploration/global issues/poverty/travel
9	8	1	2006/01/01 00:00:00.000000000	2200/01/01 00:00:00.000000000	20	gaming
10	9	1	2006/01/01 00:00:00.000000000	2200/01/01 00:00:00.000000000	21	videoblogging/web
11	10	1	2006/01/01 00:00:00.000000000	2200/01/01 00:00:00.000000000	22	blogs/humanity/people
12	11	1	2006/01/01 00:00:00.000000000	2200/01/01 00:00:00.000000000	23	comedy/humor
13	12	1	2006/01/01 00:00:00.000000000	2200/01/01 00:00:00.000000000	24	entertainment
14	13	1	2006/01/01 00:00:00.000000000	2200/01/01 00:00:00.000000000	25	communication/news/politics
15	14	1	2006/01/01 00:00:00.000000000	2200/01/01 00:00:00.000000000	26	architecture/art/creativity/fashion/howto/style
16	15	1	2006/01/01 00:00:00.000000000	2200/01/01 00:00:00.000000000	27	education
17	16	1	2006/01/01 00:00:00.000000000	2200/01/01 00:00:00.000000000	28	algorithm/anthropology/computers/neuroscience/science/software/technology
18	17	1	2006/01/01 00:00:00.000000000	2200/01/01 00:00:00.000000000	29	activism/business/nonprofits
19	18	1	2006/01/01 00:00:00.000000000	2200/01/01 00:00:00.000000000	30	movies
20	19	1	2006/01/01 00:00:00.000000000	2200/01/01 00:00:00.000000000	31	anime

5. Video Dimension

The Video Dimension contains all the videos that were posted on Youtube from 2006-2018. We assigned the 'Video_dim_id' as the surrogate key on Pentaho and the native key as the 'video_id'. We embellished it by adding the description of the what the video is about and also the channel that the video was posted from. The type 2 slowly changing dimension has been set up to capture any changes in the source system.



Execution Results

Execution History | Logging | Step Metrics | Performance Graph | Metrics | Preview data

#	Stepname	Copynr	Read	Written	Input	Output	Updated	Rejected	Errors	Active	Time	Speed (r/s)	input/output
1	Video CSV file input	0	0	40949	40950	0	0	0	0	Finished	7mn 23s	92	-
2	Video Dimension lookup/update	0	40949	40949	11865	8255	0	0	0	Finished	9mn 11s	74	-

video_dim_id	version	date_from	date_to	video_id	title	channel_title	thumbnail_link	description
0	1	<null>	<null>	<null>	<null>	<null>	<null>	<null>
1	1	2006/01/01 ...	2019/05/11 ...	2kyS6SvSYSE	WE WANT TO TALK ABOUT OUR MARR...	CaseyNeistat	https://i.ytimg.com/vi/2kyS6SvSYS...	SHANTELL'S CHANNEL - https://www.y...
2	1	2006/01/01 ...	2200/01/01 ...	1ZAPwfrtAFY	The Trump Presidency: Last Week Tonig...	LastWeekTonight	https://i.ytimg.com/vi/1ZAPwfrtAFY...	One year after the presidential election, ...
3	1	2006/01/01 ...	2200/01/01 ...	5qpjK5DgCt4	Racist Superman | Rudy Mancuso, King...	Rudy Mancuso	https://i.ytimg.com/vi/5qpjK5DgCt...	WATCH MY PREVIOUS VIDEO â–¶ \n\nS...
4	1	2006/01/01 ...	2200/01/01 ...	puqaWrEC7tY	Nickelback Lyrics: Real or Fake?	Good Mythical M...	https://i.ytimg.com/vi/puqaWrEC7t...	Today we find out if Link is a Nickelback...
5	1	2006/01/01 ...	2200/01/01 ...	d380meD0W0M	I Dare You: GOING BALD!?	nigahiga	https://i.ytimg.com/vi/d380meD0...	I know it's been a while since we did thi...
6	1	2006/01/01 ...	2019/05/11 ...	gHZ1Qz0KiKM	2 Weeks with iPhone X	iJustine	https://i.ytimg.com/vi/gHZ1Qz0KiK...	Using the iPhone for the past two week...
7	1	2006/01/01 ...	2200/01/01 ...	39idVpFF7NQ	Roy Moore & Jeff Sessions Cold Open ...	Saturday Night Live	https://i.ytimg.com/vi/39idVpFF7N...	Embattled Alabama Senate candidate R...
8	1	2006/01/01 ...	2200/01/01 ...	nc99ccSXST0	5 Ice Cream Gadgets put to the Test	CrazyRussianHack...	https://i.ytimg.com/vi/nc99ccSXST0...	Ice Cream Pint Combination Lock - http:...
9	1	2006/01/01 ...	2200/01/01 ...	jr9QtXwC9vc	The Greatest Showman | Official Trailer ...	20th Century Fox	https://i.ytimg.com/vi/jr9QtXwC9vc...	Inspired by the imagination of P.T. Barn...
10	1	2006/01/01 ...	2200/01/01 ...	TUmyygCMMGA	Why the rise of the robots wonâ€™t m...	Vox	https://i.ytimg.com/vi/TUmyygCM...	For now, at least, we have better things ...
11	1	2006/01/01 ...	2200/01/01 ...	9wRQljFNDW8	Dion Lewis' 103-Yd Kick Return TD vs. ...	NFL	https://i.ytimg.com/vi/9wRQljFND...	New England Patriots returner Dion Lewi...
12	1	2006/01/01 ...	2200/01/01 ...	VifQlJit6A0	(SPOILERS) 'Shiva Saves the Day' Talke...	amc	https://i.ytimg.com/vi/VifQlJit6A0/...	Shiva arrives just in time as King Ezekiel ...
13	1	2006/01/01 ...	2200/01/01 ...	5E4ZBSInqUU	Marshmello - Blocks (Official Music Vid...	marshmello	https://i.ytimg.com/vi/5E4ZBSInqU...	WATCH SILENCE MUSIC VIDEO â–¶ http...
14	1	2006/01/01 ...	2200/01/01 ...	GgVmn66oK_A	Which Countries Are About To Collapse?	NowThis World	https://i.ytimg.com/vi/GgVmn66oK...	The world at large is improving, but so...
15	1	2006/01/01 ...	2019/05/11 ...	TaTleo4cOs8	SHOPPING FOR NEW FISH!!!	The king of DIY	https://i.ytimg.com/vi/TaTleo4cOs8...	Today we go shopping for new fish for ...

6. TED Talk Fact Table

Since the dimension table transformations are stored in the target database, we could now create fact tables. By using the dimension lookup function, we are able to connect the Speaker, Talk, Topic, and Date dimension for the TED Talk Fact Table. The dimensions that connect to the YouTube Fact Table are Video, Topic, and Date dimensions. To generate percentages and to insert values such as comment, dislikes, likes and views for each of the fact table we had to import the 'calculation' step. This allowed us to gain an insight on engagement rate of how many people were liking versus the total number of views. The percentages for TED Talk Fact table were way below zero. While we were able to get whole number percentages for YouTube Fact Table. This shows us that there was considerable amount of people that were posting more comments on video in comparison to TED Talk.

Execution Results

Execution History | Logging | Step Metrics | Performance Graph | Metrics | Preview data

#	Stepname	Copynr	Read	Written	Input	Output	Updated	Rejected	Errors	Active	Time	Speed (r/s)	input/output
1	TED CSV file input	0	0	2374	2375	0	0	0	0	Finished	0.0s	49,479	-
2	Speaker Dim lookup	0	2374	2374	2374	0	0	0	0	Finished	37.1s	64	-
3	Talk Dim lookup	0	2374	2374	2374	0	0	0	0	Finished	38.5s	62	-
4	Topic Dim lookup	0	2374	2374	2374	0	0	0	0	Finished	41.6s	57	-
5	Date Dim lookup	0	2374	2374	2374	7	0	0	0	Finished	41.7s	57	-
6	ComentOverView	0	2374	2374	0	0	0	0	0	Finished	41.7s	57	-
7	Load Ted Fact Table	0	2374	2374	0	2374	0	0	0	Finished	41.7s	57	-

Ted_Deg_id	Speaker_dim_id	Talk_dim_id	Topic_dim_id	date_dim_id	Coments	Views
1	1	1	14	4761	4553.0	47227110
2	2	2	2	4761	265.0	3200520
3	3	3	12	4762	124.0	1636292
4	4	4	17	1	200.0	1697550
5	5	5	7	4763	593.0	12005869
6	6	6	27	4764	672.0	20685401
7	7	7	11	4762	919.0	3769987
8	8	8	14	4765	46.0	967741
9	9	9	27	4764	852.0	2567958
10	10	10	27	4761	900.0	3095993
11	11	11	14	1	79.0	1211416
12	12	12	14	1	55.0	387877
13	13	13	7	4765	71.0	693341
14	14	14	16	4766	242.0	4531020
15	15	15	15	4765	99.0	358304
16	16	16	3	4765	325.0	2702470
17	17	17	7	4762	88.0	1415724
18	18	18	17	4765	108.0	966439

7. YouTube Fact Table

Execution Results

Execution History | Logging | Step Metrics | Performance Graph | Metrics | Preview data

#	Stepname	Copynr	Read	Written	Input	Output	Updated	Rejected	Errors	Active	Time	Speed (r/s)	input/output
1	YouTube CSV file input	0	0	40949	40950	0	0	0	0	Finished	42mn 19s	16	-
2	Video Dim lookup	0	40949	40949	40949	102	0	0	0	Finished	55mn 53s	12	-
3	Topic Dim lookup	0	40949	40949	40949	0	0	0	0	Finished	55mn 53s	12	-
4	Date Dim lookup	0	40949	40949	40949	0	0	0	0	Finished	55mn 53s	12	-
5	LikesOverViews	0	40949	40949	0	0	0	0	0	Finished	55mn 53s	12	-
6	Table output	0	40949	40949	0	40949	0	0	0	Finished	55mn 53s	12	-

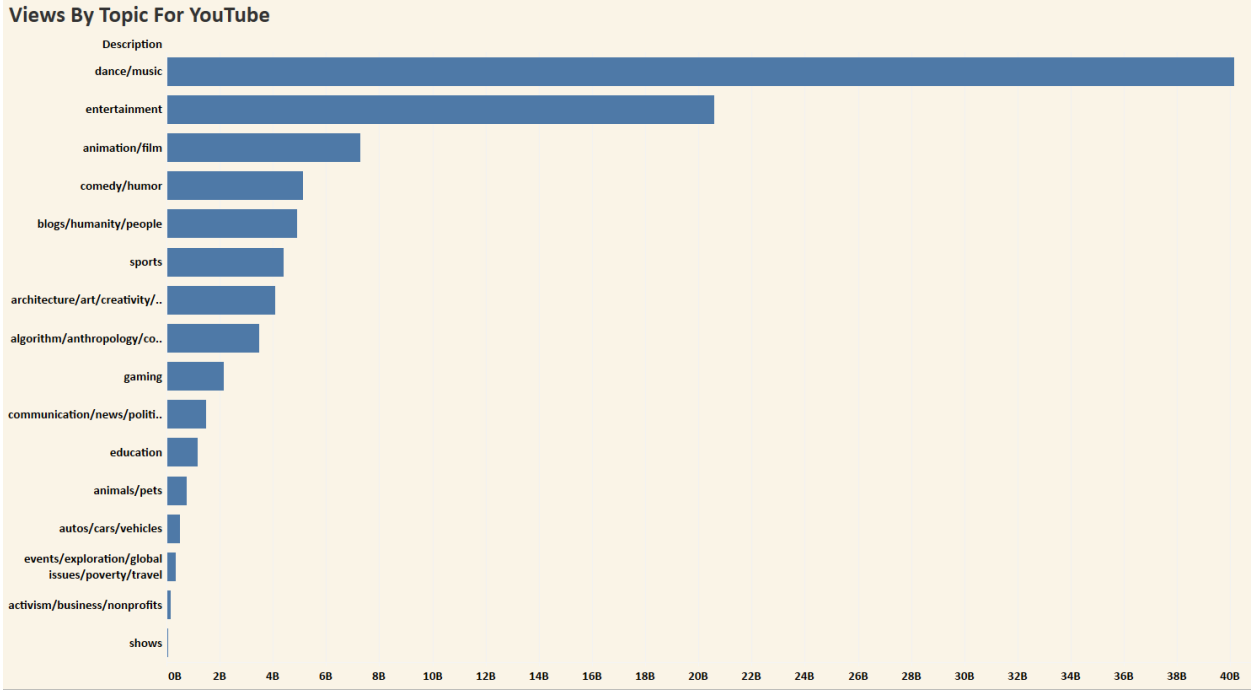
YouTube_Deg_id	video_dim_id	Topic_dim_id	date_dim_id	percentage_likes	Likes	Dislikes	Coments	Views
1	213	10	4279	7.0	57527	2966	15954.0	748374
2	2	12	4279	4.0	97185	6146	12703.0	2418783
3	3	11	4278	4.0	146033	5339	8181.0	3191434
4	4	12	4279	2.0	10172	666	2146.0	343168
5	5	12	4278	6.0	132235	1989	17518.0	2095731
6	510	16	4279	8.0	9763	511	1434.0	119180
7	7	12	4278	0.0	15993	2445	1970.0	2103417
8	8	16	4278	2.0	23663	778	3432.0	817732
9	9	1	4279	0.0	3543	119	340.0	826059
10	10	13	4279	4.0	12654	1363	2368.0	256426
11	11	5	4279	0.0	655	25	177.0	81377
12	12	12	4279	1.0	1576	303	1279.0	104578
13	13	3	4279	16.0	114188	1333	8371.0	687582
14	14	13	4278	1.0	7848	1171	3981.0	544770
15	243	4	4278	3.0	7473	246	2120.0	207532
16	16	16	4279	12.0	9419	52	1230.0	75752
17	247	11	4278	2.0	8011	638	1256.0	295639
18	18	15	4279	6.0	5398	53	385.0	78044

IV. Business Intelligence and Application Design

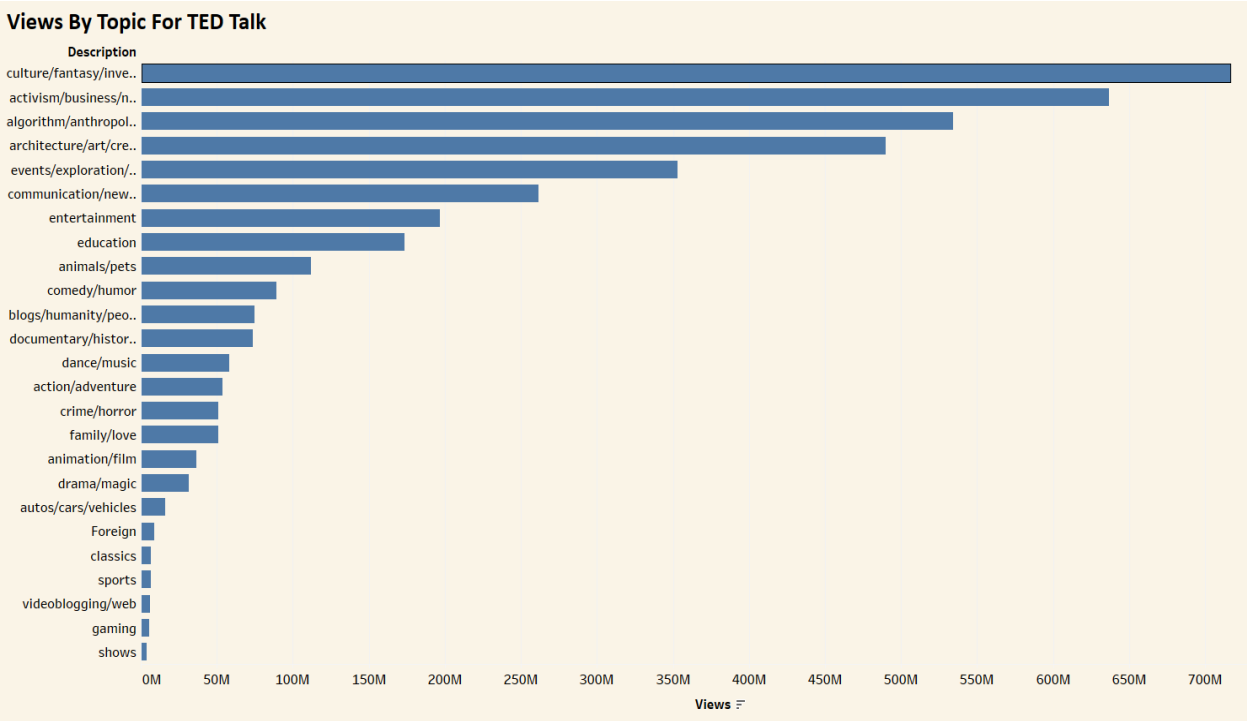
This part consists of presenting different type of analysis we can perform to extract knowledges from the data. The comparative analysis, the trending analysis, the contribution analysis, and the correlation analysis will be presented in this section.

1. Business Comparative Analysis

Comparative analysis allows us to compare the total number of views based on each topic. When we performed our comparative analysis on the Description and the Number of Views metrics, we found that “Dance/Music” was the clear most-watched topic and was ranked number 1 on YouTube. Using this type of analysis, gave us an idea of what topics people are more interested to watch on YouTube compare to other platforms. To maintain this number of views, YouTube should keep up promoting and posting more of such type of videos.

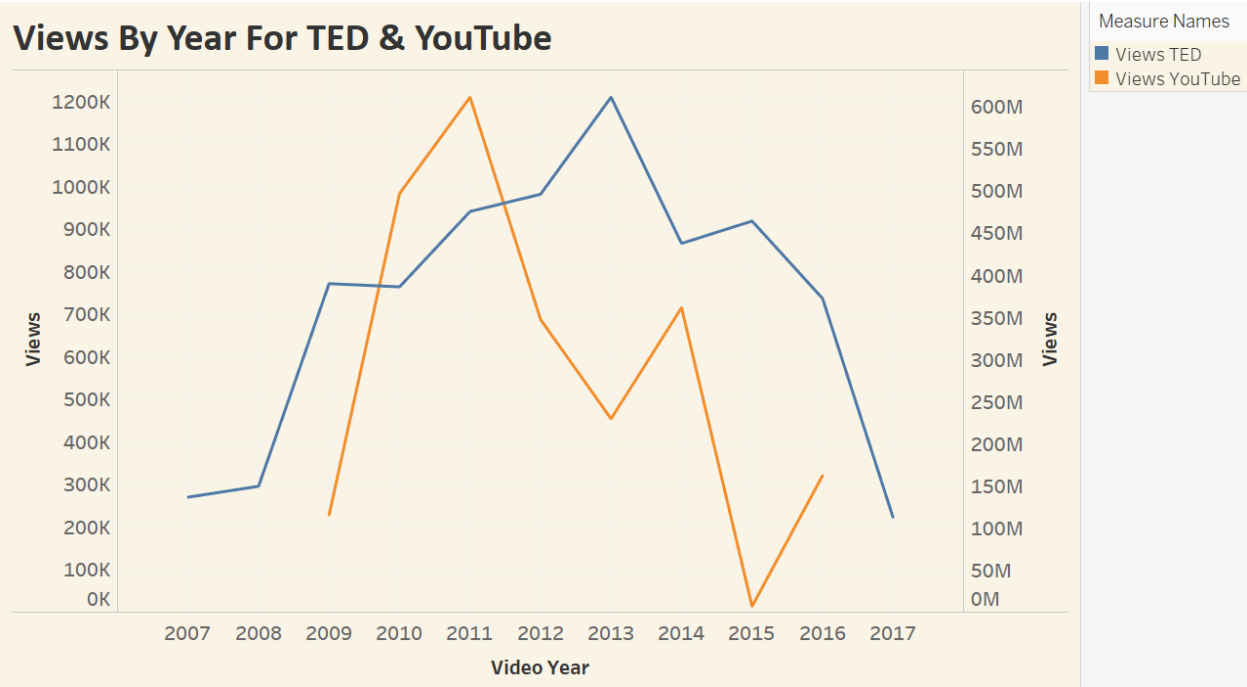


We then did another comparative analysis on the Ted Talks with the same topics where “culture/fantasy/invention” and “activism/business/nonprofits” were on the top of the lists respectively. This shows that more audiences prefer to watch Cultural, Fantasy, and Business-related videos on Ted. Thus, to increase the number of views in the site, Ted needs to have more videos that talk about those topics.



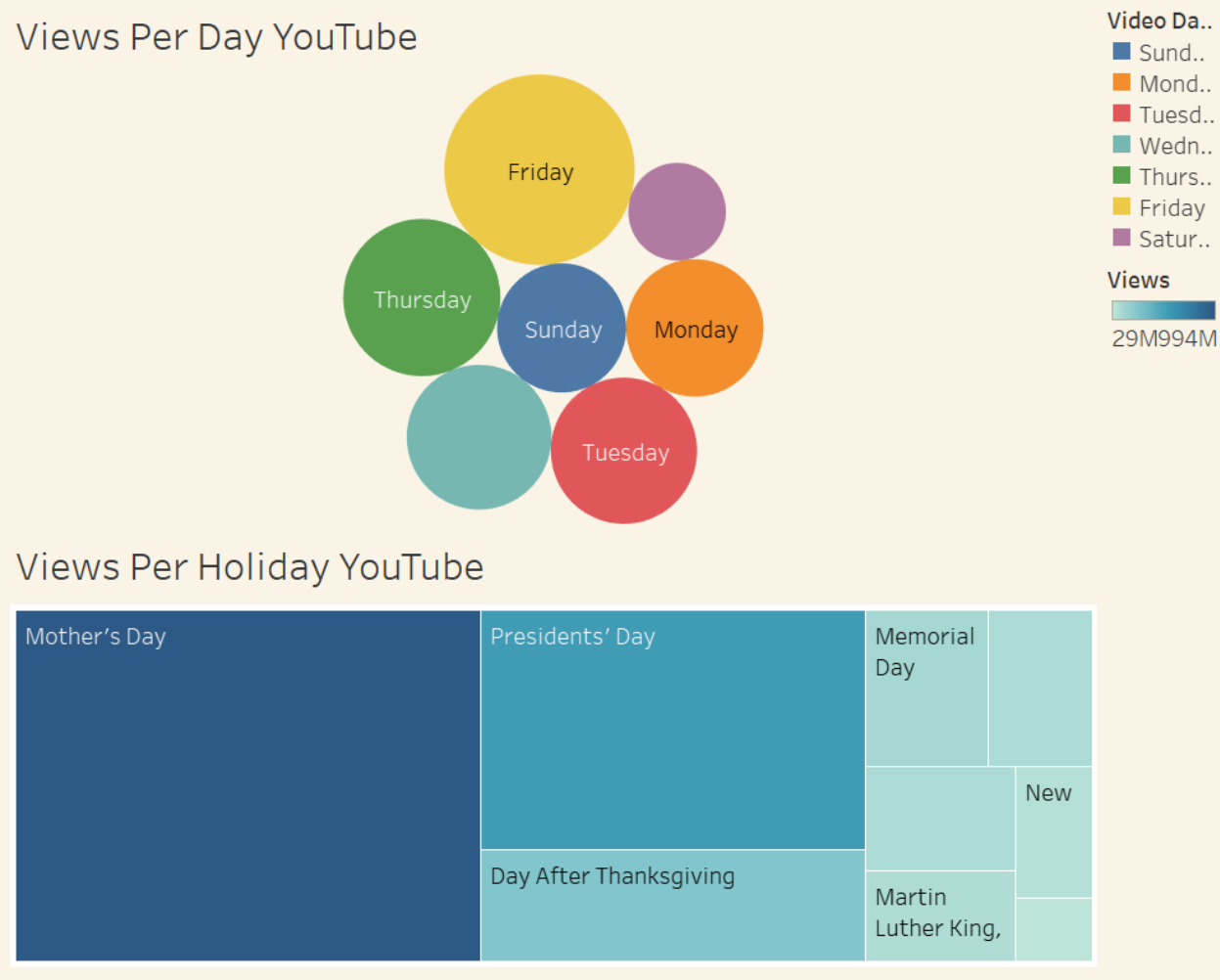
2. Time Series Analysis

Using the Trend Analysis, it allows us to see the changes that have taken place over the years. We were able to compare the number of views both in TED and YouTube from late 2000ths to 2017. YouTube, whose measures are in millions, experienced extreme increases and decreases over that period. TED, on the other hand, had a gradual increase in the number of views from beginning of the period until its peak in 2013. Then it experienced a downward slope that continued to 2017. We can also notice an inverse relationship between the number of views of YouTube and TED in the interval 2011 to 2014 and 2015 to 2017. The correlation analysis will give us more details about this inverse relationship that we spotted.

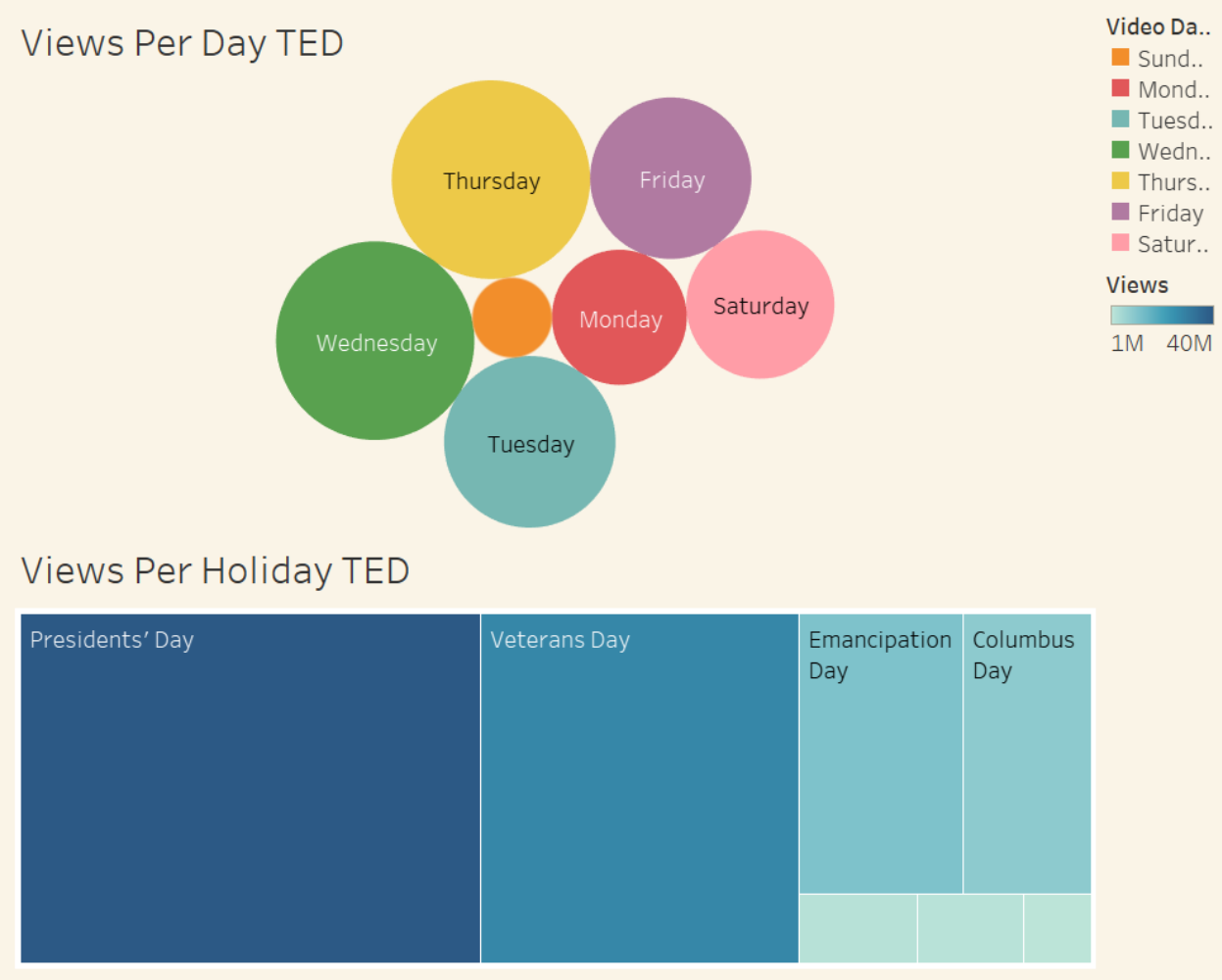


3. Contribution Analysis

Contribution Analysis helped us understand exactly what day of the week the most-watched YouTube video was. As illustrated in the bubble chart above, Friday had the highest views. As Friday is the last day of the work week for most people, it is the day that they turned to YouTube to listen to their music while going out for drinks and dinners. Using the same type of analysis on views per holiday, Mother's Day and President's Day had the most views compared to other holidays. This suggest that on Mother's Day, viewers watched more YouTube videos with families and relatives.

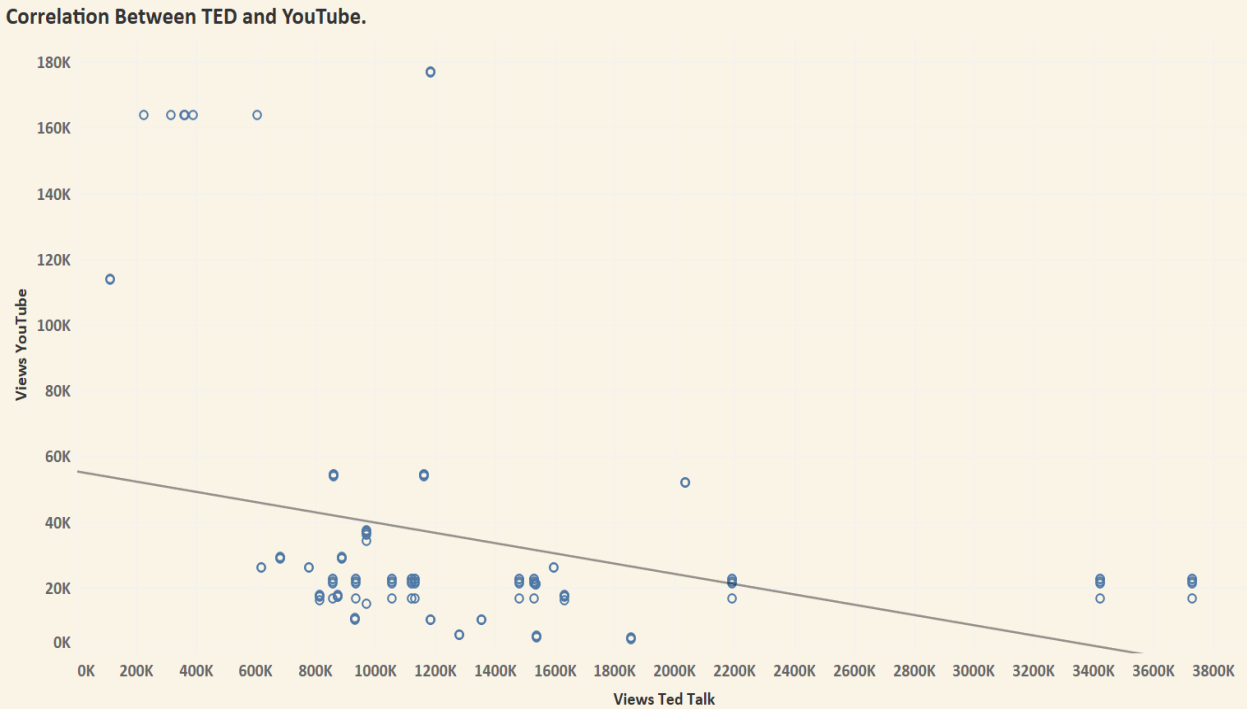


TED talks and videos had higher views on Thursdays and Wednesdays that were more spread out than the views on YouTube. The highest-ranking topics was among Culture, Activism and Business that motivated the TED audiences to watch on different days. President's Day and Veterans Day was the popular days among the holidays. By using this analysis, we can conclude that the top ranked topics on TED were positively related to the number of views on the higher viewer holidays.



4. Correlation Analysis

The number of views of YouTube Videos tends to decrease as the number of views of TED Talks increases. It means that both websites are not complementary. People tend to substitute one for the other. That is the reason why the analysis of the number of views per day shows that people mostly visit TED on Wednesday and Thursday and migrate to YouTube on Friday. In addition, people spend more time on YouTube on Mother's day and switch to TED during the Presidents' day.



Conclusion

This project aimed to demonstrate that we can apply tools we covered in class in a real project. We followed Kimball's methodology to build an Enterprise Data Warehouse hosted in the cloud. Throughout the project we faced many difficulties, one of the issues we faced with was deciding on a target database and being able to connect Pentaho Data Integration tool to the target database since we needed to install the right driver in both Pentaho and the data integration folder. After installing the driver, the second issue was about the IP address of the SQL Server database in AZURE that didn't match with the one of our computer. The easiest part was after running the ETL, working with Tableau for analysis. We also faced various technical problems and running into them helped in better being able to understand how to work with these tools. If we were to do this all over again, one thing we would do differently is being able to integrate more data reflect our project to gain a deeper and better insight about why and how YouTube is able to gain such a huge audience. We could have achieved this by finding a dataset that contained behavioral data about the use of internet or video-sharing websites around the world.