BIG DATA FOR DATA JOURNALISM, ENHANCED BUSINESS ANALYTICS AND VIDEO RECOMMENDATION AT GLOBO

M. Souza, J. Castellani, D. Monteiro, C.O. Queiroz

TV Globo, Brazil

ABSTRACT

Big data has become mainstream with the recent evolution of cloud infrastructures, data gathering and intelligence algorithms. Recently, TV Globo, the largest free-to-air broadcaster in Brazil, has implemented a multi-tenant big data project involving several fields of expertise, including data journalism, enhanced business analytics and video recommendation. The underlying technological architecture was built on top of several big data cloud services and based upon open-source big data frameworks. This allowed us to focus more attention on applications and less on infrastructure administration. The use of structured and un-structured data brings a lot of complexity but also achieves great results in terms of: efficiency for producing news stories, more complex analysis of key performance indicators for digital products (including multi-platform TV) and greater engagement in our video products, measured by social activity and conversation around our brand and programmes.

INTRODUCTION

For a long time, research and analytics to support the broadcast TV business has relied upon research institutes using viewer panels to determine what customers want. Big data tools allow us to change perspective: to observe how viewers behave and to *infer* what they want, without having the expenditure of asking them [1].

The advance of software virtualisation, which allows the extraction, storage, processing and analysis of diverse data, in real-time and with fine granularity at low cost, is significantly influencing the opinions of company executives. For the past few years, TV Globo has been working together with its digital partner, Globo.com [2], to become a data-driven analytics organisation.

This strategy is not confined to our technological activities, it advises our fundamental company objectives by answering questions throughout the business. The operating model we have chosen for TV Globo is the hub-and-spoke, where one area is our centre of excellence on data science, and where all other areas have their *data citizens*, trained on data analytics tools [3]. Inside TV Globo, news, research and digital media divisions have been the initial areas to engage in these activities.

In this paper, we shall first explore the big data architectural solution used by our news division: to access diverse governmental information, to generate leads for our newscasts and to support our digital news products. What we call *data journalism*.

Second, we'll present the same architectural solution but used for enhanced analytics and self-service business intelligence. Here our digital media and research division have been monitoring the most important business performance indicators of Globo digital products and experimenting with multi-platform TV analysis.

Third, because broadcast TV content has driven social conversation for the past 50 years in Brazil, we have implemented a dynamic and real-time video recommendation tool, with a machine learning algorithm. This will attempt to infer what video scenes from TV Globo's free-to-air broadcasts are trending on social networks, and then to recommend VOD assets for our Globo Play OTT product.

Finally, some conclusions, lessons we have learned, and a look at our next steps.

DATA JOURNALISM

Data journalism is the ability to explore and combine different sources of data into relevant news facts. The data journalism procedure, which has become a core activity in every significant newsroom across the world, follows different approaches: data can either be treated as the single reference, or it can be the tool with which the plot is explored, or it can be both.

Although the term is used in different ways and the definition may vary between different newsrooms, the goal of the data journalism process is to reveal a story that is relevant and was unknown before data mining was carried out. To achieve this goal, changes had to be made: it was necessary to adapt the profile of the newsroom team. Technology became a key expertise in the journalism field, and as a result of the changes, journalists, designers and developers now share the same newsroom space [4].

One of the recent technology initiatives on data journalism in Brazil was to make all public data available to newsrooms for the production of news stories, including information from: federal, state and city social and economic resources; the Environmental Ministry, IBAMA [5]; the Public Healthcare Programme, DATASUS [6]; the Electoral Superior Court, TSE [7]; the Brazilian Institute of Geographical, Social & Economic Statistics, IBGE [8]; and the Public Security Secretariat of Rio de Janeiro, SESEG [9]. Figure 1 shows the system architecture and Figure 2 shows the solution architecture proposed by the Globo technology team.

This proved to be a great challenge because we had to deal with structured, semi-structured and unstructured data and in many different formats: html, csv, xls, shp, dbf e pdf. We used Amazon EC2 cloud processing [10] to handle the ingestion processes from the different public sources.

All files with raw data were ingested into our *data lake* [11] hosted on Amazon AWS S3 cloud storage [12]. Besides the outside data, we also ingested some private in-house data from our Avid iNews platform.

Using Amazon EMR [13] (a cloud Elastic Map Reduce solution), within an Apache Hadoop data processing framework [14] for the data set, we run Spark, which is a fast and general computing engine for Hadoop data [15], and prepare data for future analysis. We chose PARQUET as the output format for the processed data because of its compressed, efficient columnar data representation which is available to any product in the Hadoop ecosystem [16]. This structured data was then ingested back again into a different *bucket* on Amazon S3, what we call a *data reservoir* in Figure 1.

Having the data ready for analysis was the first part. For the analysis itself, we chose the Amazon AWS Athena tool for interactive SQL (Structured Query Language) queries on our data reservoir [17].

Instead of centralising all the knowledge on accessing and analysing the data, we decided to train the data journalists on the Athena tool and taught them how to run simple-to-complex questions on the database, in the manner of our hub-and-spoke model, where technology is the centre of excellence and data-knowledgeable staff are spread across our business areas. Figure 3 shows a screenshot of the Athena query dashboard.

Besides SQL queries, we made available a Tableau application to access the columnar database for building a dashboard visualization that could be directly published on our news portal, called G1 [18]. For even deeper analysis our technology group is now starting to use R-Studio [19] and Jupiter notebook [20].

To evaluate efficiency of this activity, we processed the same data found on a study published by G1 [21]. Using our original research methods involving data research and manipulation on Microsoft Excel spreadsheets, it took 1 week to answer a typical business question. Using this new big data environment, it took 7 seconds to process a typical query.

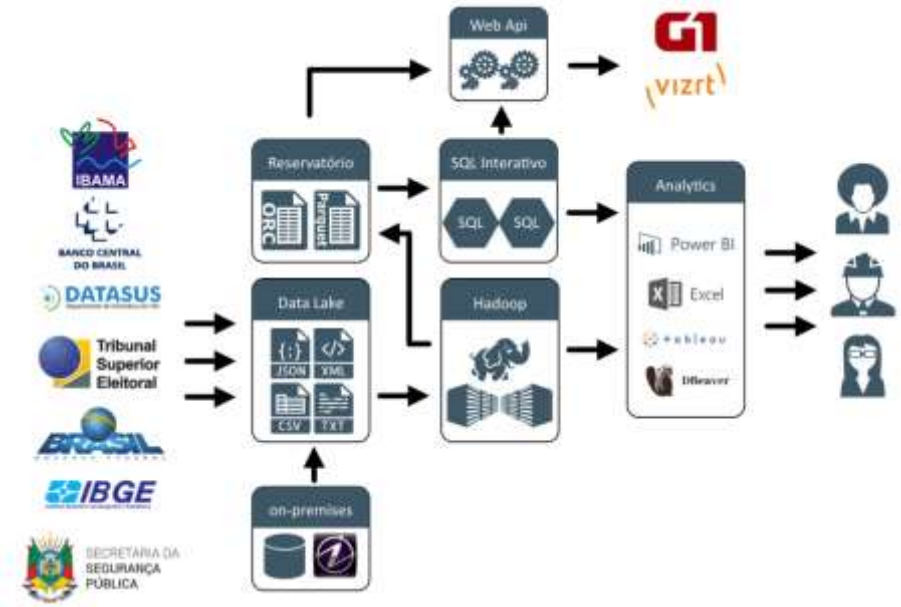

Figure 1 - System architecture for data journalism

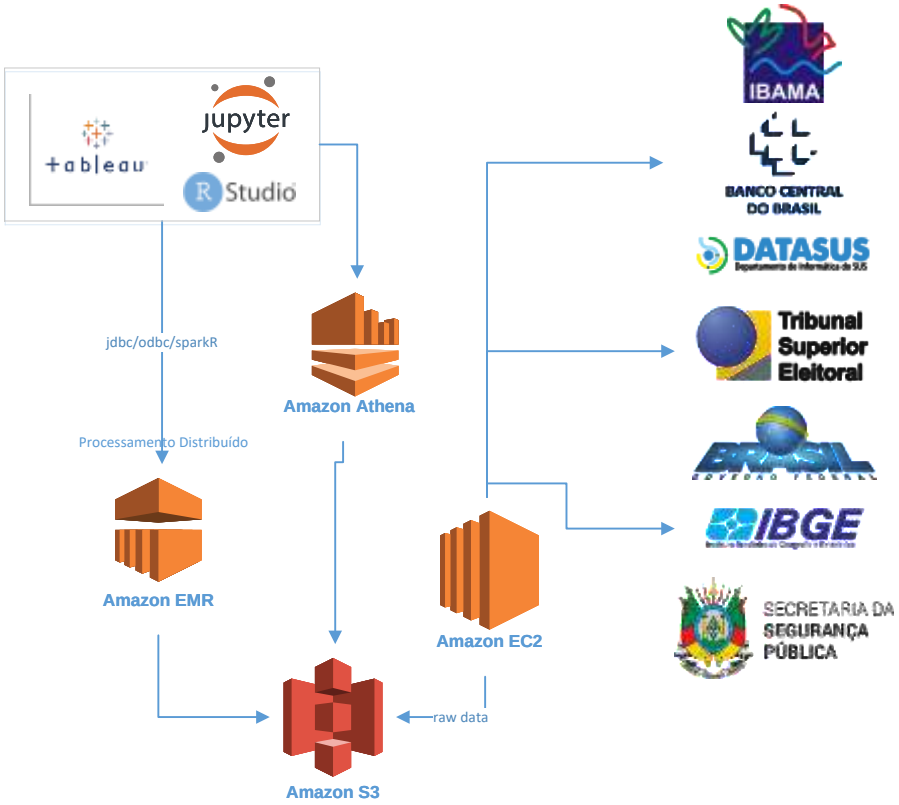

Figure 2 - Solution architecture for data journalism

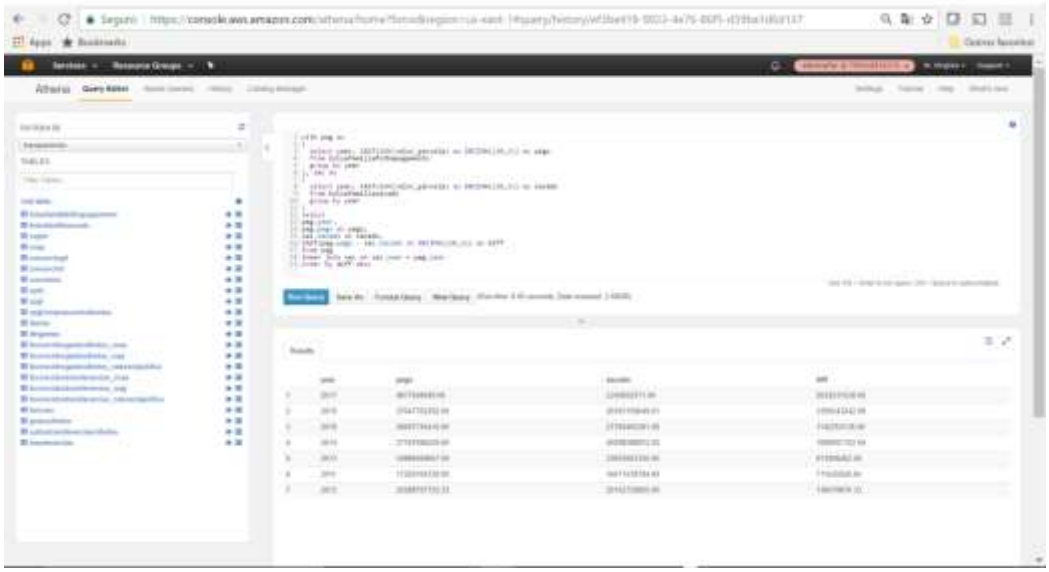

Figure 3 - Athena dashboard for SQL interactive queries

ENHANCED BUSINESS ANALYTICS

Business analytics is the use of data analysis to increase the viability and efficiency of handling the enormous number of variables that directly or indirectly affect the business. The key steps to the process are: (i) Define which data are important to the business; (ii) Determine different ways of generating this data; (iii) Analyse previous data lists representing what has already occurred; (iv) Predict and anticipate trends, behavioural responses and requirements for the future of the business; (v) Build a business plan based on the scenarios developed.

Business analytics is an umbrella term that includes data warehousing, data analysis, business information and performance management, analytic application, governance and risk. With the advance of big data tools, more data can be observed and more complex correlations can be performed [21] than ever before.

Since metrics for multi-platform TV is a trending topic in all broadcast businesses [22], we have focused our attention on metrics relating to audiences and consumption of our digital video assets across different web portals hosted by globo.com and also Globo Play. Globo Play is a TV Globo OTT product, which includes a catalogue of catch-up VOD assets, an archive library and live simulcast of local affiliates [23].

As shown in Figures 4 and 5, we used a similar system architecture as the data journalism case with some minor differences, as explained below.

For business analytics, we ingested raw metrics from Google Analytics, Comscore Media Metrix, Adobe Digital Analytics and TV audience ratings from Kantar IBOPE Media (Brazilian Institute of Public Opinion and Statistics) into our data lake, hosted on Amazon S3. TV ratings were imported manually from our existing in-house database. Digital metrics were extracted using Google Big Query from Globo's Google Analytics 360

account. For performance and cost reasons, there was an intermediary data lake on Google Cloud Services [24] which received data in json format. Amazon AWS Lambda [25] would orchestrate this process guaranteeing the synchronisation between Google Cloud Storage and Amazon AWS S3, besides the ingest of Comscore and Adobe data.

Apart from the SQL queries using Amazon AWS Athena for advanced analytics on digital and TV metrics, and correlating off and online consumption of video assets, the technology group is also making available a dashboard with KPIs (Key Performance Indicators) related to: audience ratings, engagement evolution, subscription sales, market comparison and brand awareness.

To evaluate the efficiency of the solution, we selected an important KPI which is the number of users who are tracked when navigating through Globo's websites. We ran a query that uses, not only audience data, but also matches this with login activities. This would previously have taken weeks and much manpower to execute.

Figure 6 show the results of a query that took seconds to analyse one month's data containing billions of events. It is interesting to note that even with the variation of total unique users, from 40,000 to 80,000 daily, the absolute number of logged-in users is stable at around 10,000. One way to interpret this is that those who are logged-in, see value in the products and have a good retention rate. Another is that we have a low percentage of logged-in unique visitors, and we need to work on this with the product teams.

Here is a list of KPIs that we have been working on: reach by number of unique users on Globo Play vs TV ratings reach; total video hours consumed by users; entry points (referral) for Globo video consumption; video views by users vs total video hours consumed by users vs device vs connection type; lifetime value of a subscriber; social networks mentions of programs vs total number of video views; average of video sharing vs number of subscription sales; conversion of subscription trials vs TV ratings.

For our next steps, as other broadcast companies have tried [26], we are starting an experiment called Total Audience. The idea is to try to correlate TV and digital metrics and achieve the *Holy Grail* of multi-platform advertising intelligence.

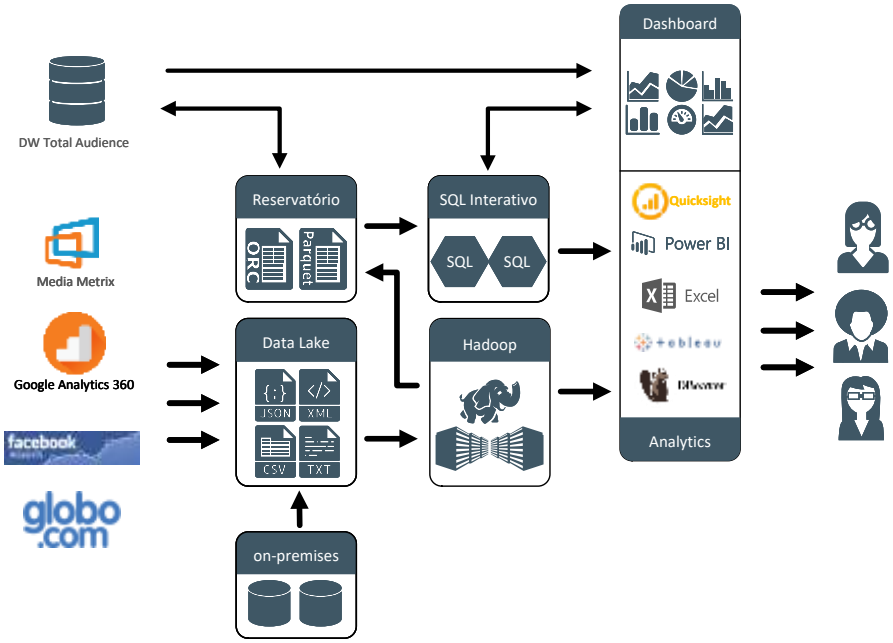

Figure 4 - System architecture for digital analytics

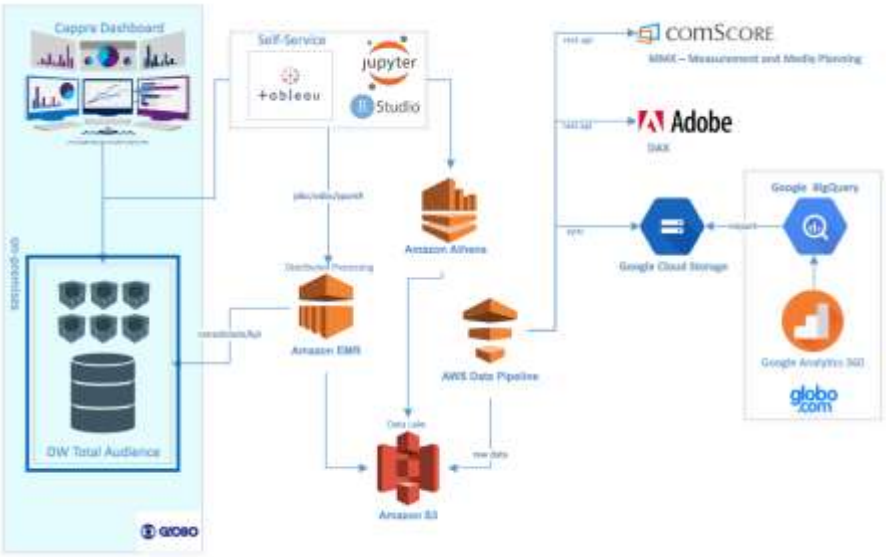

Figure 5 - Solution architecture for enhanced business analytics

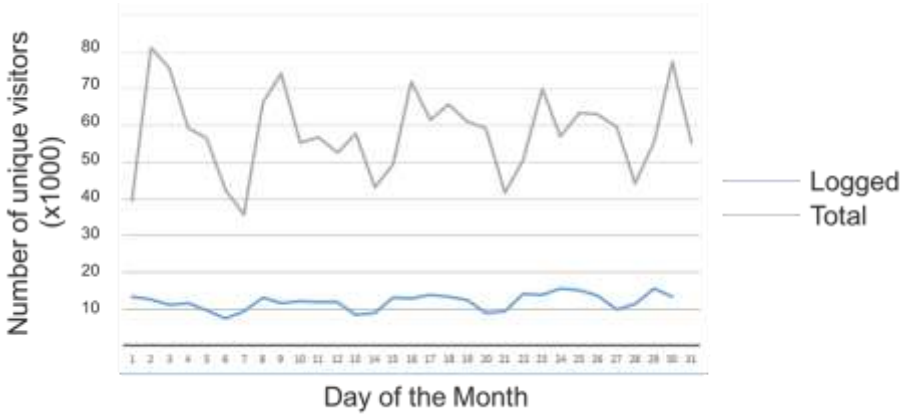

Figure 6 - Number of logged-in unique visitors vs total unique visitors on Globo digital properties

VIDEO RECOMMENDATION

Because broadcast TV content in Brazil has millions of primetime viewers, this generates many social repercussions, and in today's world, this is directly reflected on social media. Since Globo also runs an OTT product, we have developed a dynamic, real-time video recommendation tool, to suggest videos to Globo Play OTT users from our VOD catalogue. The recommendations match social conversation captured by Oracle's Endeca solution. Figure 7 shows the system architecture that we have developed.

This tool tries to infer the most relevant content from the most popular discussion topic, by using techniques from search engines and natural language processing [27] [28]. The system reads the last hours of social media comments and video publishing synopses, and connects them by their keywords. These are extracted using a TF-IDF (Term Frequency - Inverse Document Frequency) score of each of the terms in the documents [29]. In summary, the relevance of a term is proportional to its presence in a document and its rarity in the document collection. The score of similarity of two texts can then be calculated simply using the cosine distance between the feature vectors.

Once there is a relevance score between a social comment and a video description, it is possible to create a ranked list of videos according to this score. This way, a content offering can be driven not only by the consumption of the user but also by a broader context: the current buzz of the web.

The resulting recommendations have been proving very accurate. We have compared Google Analytics video views measurement of videos that are recommended based on frequency for the past hour ("[Home] Trilhos Mais Vistos") with videos recommended using this new tool based on machine learning of social conversation ("[Home] Trilhos Mais Falados"), Figure 8.

The graph in Figure 9 shows the number of users that have played at least one video from within the referred area. There was a 35% increase of video-views among unique users when using the new tool.

More work continues to be done in order to measure its performance and improve it even further with other natural language processing techniques.

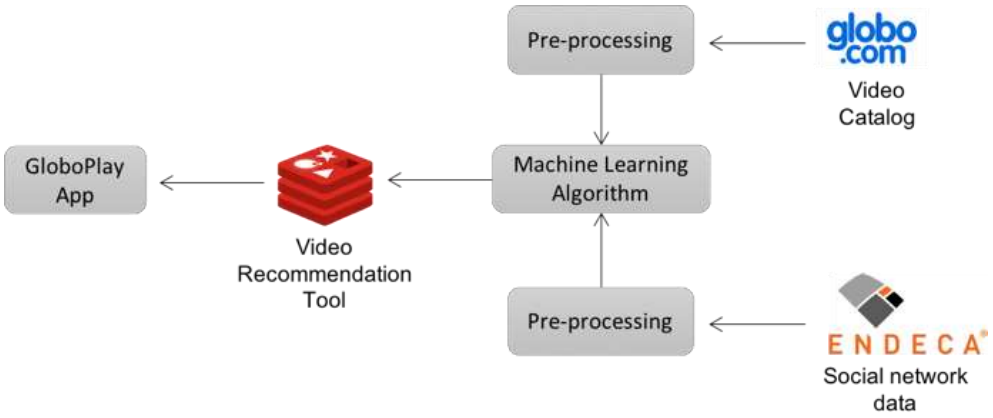

Figure 7 - System architecture for "Mais Falados" video recommendation tool

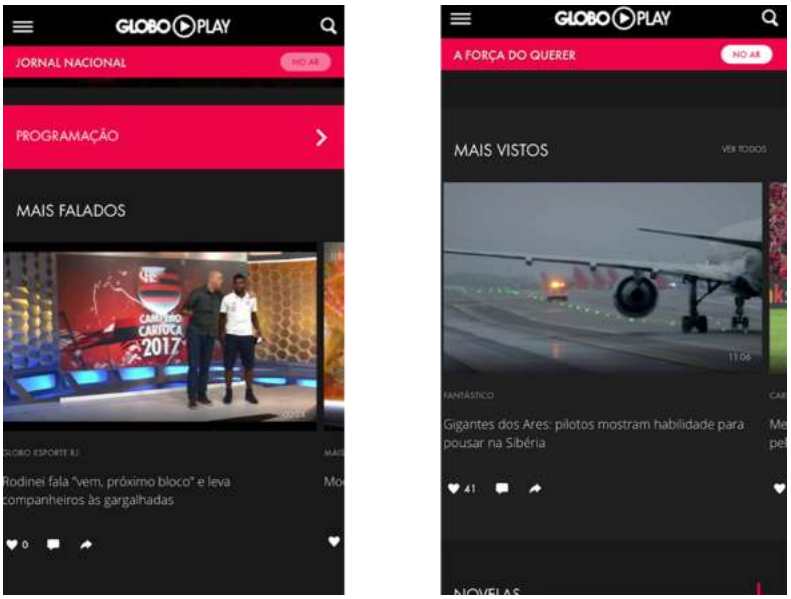

Figure 8 - "Mais Falados" area (recommendation tool) and "Mais Vistos" area (frequency of views)

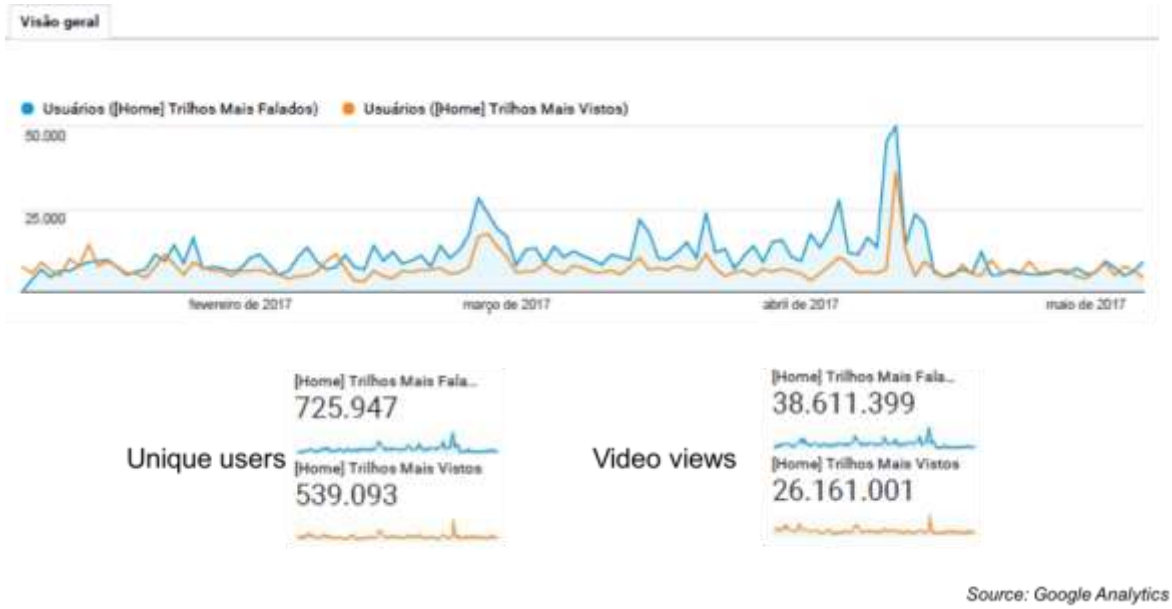

Figure 9 - Increase in video consumption using the new recommendation tool based on machine learning algorithms

CONCLUSIONS

News editors are now running analyses with complex queries that would previously have taken weeks and are preparing Globo for the data journalism era. Instead of wasting time on less cognitive workloads, journalists now can spend more time on editorial tasks. For one specific research activity, we were able to reduce data processing from weeks to seconds.

Our business analytics have moved from dashboards to a more self-service and deep statistical approach. Comparing different metrics on different scales, has allowed us to make product, content and technology decisions much more quickly. We are now able to process billions of pieces of information in seconds, at low cost and with little man-effort, providing product owners with the results of complex analytics.

Also, our roadmap for Globo Play includes the priority development of new features not only on subscription impact, but on KPIs representing audience, engagement, brand impact and innovation. Our next step is to develop more *prescriptive* approach rather than merely a *predictive* one.

Recently, our technology team has implemented a natural language processing algorithm on public social networks to automatically discover which video clips of a show should be highlighted on the main screen of the application. Videos recommended using our new machine learning tool achieved a more than 30% increase in viewing, compared to videos recommended by frequency-usage statistics alone. Our next steps are to refine our recommendations based on more inputs and not only social media sources.

All three applications we have described have generated knowledge inside Globo, and most important, not only in our technology group. They all brought great efficiency to those processes which need to interpret lots of information in order to answer complex business questions. Working with such new frameworks requires skills that are yet to be developed on a mass scale within the organisation. The data-driven culture has to be fully embraced for a total digital transformation and Globo is on top of this.

REFERENCES

[1] Gartner Says It's Not Just About Big Data; It's What You Do With It: Welcome to the Algorithmic Economy http://www.gartner.com/newsroom/id/3142917

[2] BigData na Globo.com http://grandesdados.com/post/bigdata-na-globocom/

[3] Building an Analytics-Driven - Organization Organizing, Governing, Sourcing and Growing Analytics Capabilities in CPG https://www.accenture.com/dk-en/~/media/Accenture/Conversion-Assets/DotCom/Documents/Global/PDF/Industries_2/Accenture-Building-Analytics-Driven-Organization.pdf

[4] Data journalism at the Guardian: what is it and how do we do it? https://www.theguardian.com/news/datablog/2011/jul/28/data-journalism

[5] IBAMA http://www.ibama.gov.br/

[6] DATASUS http://datasus.saude.gov.br/

[7] Tribunal Superior Eleitoral http://www.tse.jus.br/

[8] IBGE http://www.ibge.gov.br/home/

[9] Secretaria de Segurança Públca Rio de Janeiro http://www.rj.gov.br/web/seseg

[10] Amazon AWS Simple Cloud Hosting (EC2) https://aws.amazon.com/ec2/

[11] What is a Data Lake? https://aws.amazon.com/big-data/data-lake-on-aws/

[12] Amazon AWS Simple Storage Service (S3) https://aws.amazon.com/s3/

[13] Amazon AWS Elastic Map Reduce (EMR) https://aws.amazon.com/emr/

[14] Apache Hadoop http://hadoop.apache.org/

[15] Apache Spark http://spark.apache.org/

[16] Parquet Documentation http://parquet.apache.org/

[17] Amazon AWS Athena https://aws.amazon.com/athena/

[18] G1 News Portal
http://g1.globo.com/

[19] R-Studio https://www.rstudio.com/

[20] Jupiter Notebook http://jupyter.org/

[21] O Saldo da Bolsa Familia, G1 Website
http://g1.globo.com/economia/noticia/2014/03/em-10-anos-r-119-bilhoes-deixaram-de-ser-sacados-do-bolsa-familia.html

[22] Business analytics in the age of big data https://www.london.edu/faculty-and-research/lbsr/business-analytics-in-the-age-of-big-data#.WQ_iC4nyuSM

[23] Cross-channel advertising attribution: New insights into Multiplatform TV
https://www.accenture.com/us-en/_acnmedia/PDF-18/Accenture-New-Insights-Into-Multiplatform-TV.pdf

[24] Globo Play https://globoplay.globo.com/

[25] Google Cloud Services & Big Query https://cloud.google.com/bigquery/

[26] Amazon AWS Lambda https://aws.amazon.com/lambda/

[27] ChengXiang Zhai, Professor - University of Illinois at Urbana-Champaign, Coursera Course: Text Retrieval and Search Engines
https://www.coursera.org/learn/text-retrieval

[28] Brandon Rose - Document Clustering with python
http://brandonrose.org/clustering

[29] TF-IDF http://www.tfidf.com/