Casey Bramlett
Homework 1
2/3/2024
CS488

Case Study #1: Google

Business Case Evaluation

Google's Big Data initiative started from the need to completely change the way users searched and received information. Traditional search engines worked by keyword matching, which was insufficient for giving relevant search results. This is where Google developed new ideas like PageRank, which took a deeper look at the relationships between websites and pages on the internet to determine their priority in the search engine. This revolutionized search results compared to the current competitors and gave more accurate and refined results to users, pushing them ahead of the competition. They not only enhanced user satisfaction but also set up a framework that was ideal for advertisers by making it easier to target ads.

The investment into Big Data was justified by the revenue that would come with these ads and understanding user behavior. Google's advertising platform, AdSense, could now deliver personalized ads that were more tailored to users because of the billions of search queries and large datasets that Google was processing. This led to a clear financial push for advanced data analytics because of the increased revenue, paving the way for later, larger scale technologies from Google like MapReduce and BigQuery.

Also, Google had a commitment to innovation that went well beyond just its search engine. The business case evaluation showed that the knowledge gained from Big Data could drive other areas, such as the self driving car. These initiatives aimed at expanding into new markets and creating new revenue streams. This evaluation was very beneficial to Google in a way that paved the way for future innovations and ensured that these Big Data analytics aligned with business objectives and long term growth.

Data Identification

Google found many different internal and external datasets to support their Big Data ventures. Internal data was mostly from search logs, user queries / query history, user interaction on pages, and the metadata that came from Google's bots crawling for information on billions of web pages. Externally, Google got data from various sources and datasets like weather forecasts, historical data, financial data and travel information. By adding in data that goes across different topics and collections, they ensure the results are comprehensive and able to give very relevant information to the user.

Data Acquisition and Filtering

Google got its internal datasets from in-house systems that power the search engines and other Google services. Things like the search logs, user queries, user interactions, and other metadata is straight from Google's own data. These datasets are always being updated as more user data is collected and stored, making sure that the search engine always has the most up to date information.

In addition to the internal data, Google also brings in external data from sources like weather forecasts, historical data, ect., which they take and store details like where it came from, size, and date. This lets them keep track of it so that they can run automated checks to filter out data used in its analysis.

Data Extraction

Google's data extraction process can be guessed based on its approach of managing the extremely large amounts of information that is gathered by its bots that scrape the web. These bots constantly copy data from the billions of web pages they visit, then process key elements to build the index used in the search engine. The case study does not explicitly outline the fields, but Google must have some type of algorithm to sort the content into a structured field that can be processed by things like PageRank. This extraction process highlights Google's ability to give a user refined and relevant search results.

Data Validation and Cleansing

The data validation and cleansing process done by google is not clearly defined but this process makes sure that the large amounts of data collected by the web scraping bots is accurate and reliable. The automated checks can remove duplications, errors, random information and fit the data into the needed format. This type of quality control is critical for maintaining the standard they want to achieve.

Data Aggregation and Representation

Google aggregates data from its web crawlers and other internal systems into a structured index and powers its search engine. This aggregation combines different elements of user data into a database that can power algorithms like PageRank and semantic search. By representing the information in this organized format, Google can make sure its systems and accurately use and retrieve this data for better results.

Data Analysis

The analysis done by google consists of continuous processing of the large datasets gathered by their web crawlers, user data / information, and the other internal indexed systems. Various techniques are applied in things like MapReduce and BigQuery which can give the most relevant results. This process refines the other algorithms like PageRank by always testing and adjusting the data, prioritizing data that has better accuracy.

Data Visualization

Google's approach to data visualization can be seen by their Universal Search and Knowledge Graph. Universal search was launched in 2007 and represents diverse data into the search results. This later became the Knowledge Graph in 2012, which showed contextual information about the subject from a wide range of sources. By mixing data from a user's history, location, Google+, and Gmail, Google can create a very personalized experience that is tailored to the specific user.

Utilization of Analysis Results

Google leverages its data analysis insights to make revolutionary products. In 2012 they launched BigQuery, which allowed companies to store and process large datasets on Google's cloud platform. They effectively created a Big Data commercial service where users pay for storage and compute time. Also, Google's self-driving car project used real-time data with inputs from Google Maps to autonomously drive. They even tried to use predictive analytics to predict a flu outbreak. They continue to show how their data analysis re-invents current technologies and methodologies but also creates room for future inventions.

Case study #2 Facebook

Business Case Evaluation

To begin the Big Data analytics lifecycle, a clear business case must outline the justification and motivation for the analysis. For Facebook, the business case is justified because of its unique position as the biggest social network, with over 1.25 billion people. The users agree to give up personal information when they register such as demographics, what they are interested in, and who they are friends with. This data is then used to target very effective advertising models, giving advertisers specific access to groups like "women over 40 who love books" or "men under 25 living in the UK who love football". This has proved to drive revenue growth up significantly.

In addition to that data collections, Facebook has acquired other platforms like Instagram, WhatsApp, and Oculus Rift to stretch their ecosystem even more, further helping their business case. These moves enhance their advertising power but also give even more insight into what their users are doing and what their interests are. With the primary motive to maximize ad revenue and keep user interaction up, Facebook has very well defined their business case, opening the door for them to continue with data analysis tasks.

Data identification

For the data identification stage, Facebook targets a wide range of datasets for critical analytics. Internally, Facebook gathers user details as stated before, like demographics, interests and friends when a user signs up. They also gather interaction data like what users like, share and comment. In addition, Facebook tracks user behavior like user monitoring tools that records metrics like cursor hover times and websites users visit when leaving Facebook. Externally, data from acquired platforms like Instagram and WhatsApp give even more data for information on a user. This briefly outlines the internal and external data sources that Facebook gets its data to uncover patterns and correlations to drive its targeted advertising and other insights.

Data Acquisition and Filtering

Data Acquisition and Filtering involves gathering data from the sources, internal and externally. Internally, Facebook gets data from user profiles, shares, comments, and interactions - which continue to be gathered as the user interacts with the platform. Externally, Facebook gets data from sources such as Instagram and WhatsApp, which are other platforms that were acquired by Facebook. This automatic process is then filtered to discard unwanted or random information, a notable example being analyzing pictures to compare faces against those in a user's friend's list to suggest tagging people.

Historically, Facebook's data practices meant that once they gathered the data, it remained on their servers even if a user chose to remove their profile. But after 2010, privacy concerns grew and Facebook implemented automated filtering and storage that allowed user to permanently delete their data from the servers. This ensured that the gathered data used for analysis was relevant and followed privacy rules.

Data Extraction

For Facebook, the data extraction stage would be converting the data from its natural format into structured formats suitable for their analysis. For example, when users upload a phot, the platform scans the photo to extract facial features and probably other metadata for facial recognition. Similarly, text from posts and comments is parsed to gather key information needed for detailed analysis and for other features. This extraction process makes it so that different types of data, images, text, or interactions, can be used for analysis by being put into a structured format that Facebook's analytics systems can read and process.

Data Validation and Cleansing

The Facebook case study does not give detailed information or examples of their validation, but it is clear that there are specific algorithms and measurements that are used. This data is very diverse and unstructured, Facebook must have some way of cleaning this user data from posts to upload photos for facial recognition to avoid inaccurate predictions. For instance, data corruption and bad metadata is likely caught by processing for batch analytics and other

validation methods. Redundancy in user data, like lots of posts and other interactions can be leveraged to fill in the gaps of missing information, therefore maintaining data accuracy and supporting relevant data being passed to the other analysis stages.

Data Aggregation and Representation

Facebook has multiple sources to aggregate data from, internally and externally. Internally, there are user profiles, posts, comments, interactions and other user behavior. Externally, there is data from external sources obtained from Facebook's other acquisitions like Instagram and WhatsApp. These datasets are likely structured very differently and may intake data differently. The case study does not provide detailed specifics on how the aggregation process works, but it is clear that they create a standard view of its users and their data. This aggregation supports Facebook's core business motive by letting advertisers target certain audiences based on a wide range of concise user data.

Data Analysis

The case study talks about how Facebook's Data Science team regularly shares insights from their analysis of user behavior, which suggest that they have a systematic approach to the analysis. The case study does not give specific details of their techniques but it can be assumed that they used both confirmatory analysis to test hypotheses about user interactions and targeted ads, and exploratory analysis to try and determine new patterns in the user data. This process is crucial for better targeting strategies and making the overall experience on the platform more tailored.

Data visualization

Facebook's case study doesn't explicitly state what visualization tools are used, but it can be assumed that their data science team uses different graphs, dashboards,s and charts to make their data accessible to businesses. This visualization likely lets advertisers and other people wanting to target ads quickly see important statistics while also giving them the ability to get more detailed if needed. By taking their analytics and putting them into more visual representations, Facebook can clearly communicate insights, letting interested parties make precise decisions.

Utilization of Analysis Results

The insights that are generated from Facebook's data analysis is put to use within the platform. Based on the analysis results, Facebook's data science team likely captures new insights of user behaviors and content interactions to further their models. These models integrate into Facebook's other systems, optimizing targeted advertising and bettering the content recommendation algorithms. This way the processed analysis of the data is turned into strategies to support business decisions that will drive the platform's performance.