Assignment-8

Sai Hemanth Nirujogi
G01065588

Big Data Challenges

Data continues to grow with the enactment of the Internet of Things in every industry. With sensors everywhere around us, the data keeps overflowing every second. While a lot of value and trends are being derived from the data generated, the organizations should be able to control the data, collect it and store it otherwise there will be consequences. Some of the major challenges faced with Big Data are data privacy, real-time analytics, integration, validation, shortage of talented personnel, inadequate analytics skills of the organization, and security. [1]

Data Privacy

Market and social researcher always sought to help companies map understand and predict the user behavior until recently there are laws that kept them on the straight and narrow. But Big Data comes along as a big idea and suddenly organizations started collecting so much data that are so well and big in their breadth and depth that changed the field entirely. This is the data that every organization and government are clamoring to collect because everybody's job gets easier the more they know about everybody else. They may ask permission to originally collect it for a stated purpose then it is very likely they started using it for others. So this all comes down to predictive algorithms for marketing purposes like searching, aggregating, cross-referencing in order to generate insights. In cross-referencing datasets, Big Data analytics is putting together pieces of information in isolation as seemingly small but produces a bigger picture of troubling intelligence for everyone across all transactions, communications, and every movement. This gathered insight on an entire society lifting powerful institutions learn things about us and other wider groups make decisions about us in secret. When the algorithms find abnormal data in large datasets it can actually contribute to discrimination. For example, a 2014 white house report in big data said that web searches involving black-identifying names for example “Andre” were more likely to display as with the word arrest in them than searches with white-identifying names like “Jack”. If you don’t fit the mold or you act out of the normal to be from an undesirable grouping you will be placed into boxes that are nearly impossible to escape from. Several states in the United States are using this kind of algorithms to find the areas with more crime rate. When social infrastructure itself depends on an algorithm and that algorithm can be wrong, biased or even unfit for the purpose been seriously interfered in an arbitrary fashion big data can be helpful but it can also be dangerous now and in the future. [2]

Real-Time Analytics

We are living in a real-time world with a need for up to the minute user tracking across industries from e-commerce, security, through financial services, even web and IT analytics and monitoring. Facebook processes 2.5 billion pieces of information with 600 terabytes of data each day. Bloomberg reports 1.2 million ticks per second, Amazon reported 100 million users on Black Friday alone and Google processes 3.5 billion requests per day [3]. The number of clicks and searches happening every day is unimaginable. To calculate all these data with analytics power can break down our batch systems resulting in slow processing and slow write time with slow I/O. A heavy process of crunching the entire set of data in time and time again can lead to overload. To prevent this we can store the data and the events in a partitioned memory which can analyze massive data streams in real-time. The applications must have high storage availability and low response times which can handle massive amounts of data upto terabytes but should return insights in seconds. [4]

Data Validation

Unlike a dataset, a data source has no beginning and no end. It is either the data format may change forcing us to change the process of analytics. Data validation is a crucial and ongoing challenge which is highly complex for Organizations because they have to come up with a new technology and policy changes. It is very essential for an organization to ensure the data integrity and immutability because if the data gets modified will be of no value at all. When an erroneous information enters an organization's complex system even a minute error can lead to revenue and efficiency loss, and failure to comply with industry regulations. [5]

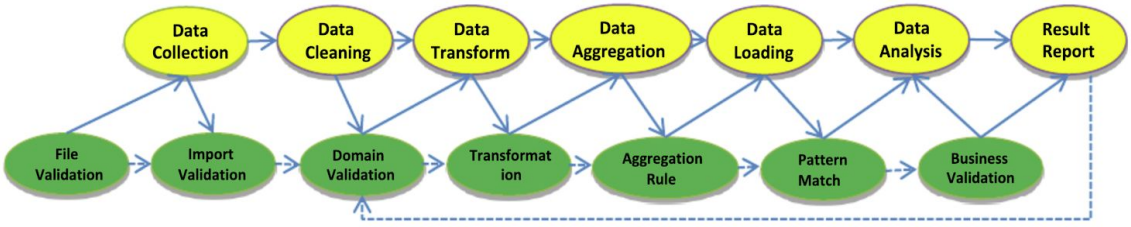

Fig.1: Multiple Data Source validation process [5]

Failed Big Data Projects

Google Flu Trends

At first, it was a paper written in Nature by a set of researchers from Google research and various academic institutions. What they are looking at was whether they could take search queries like fever, aching, joints, headache. By taking the search queries from across the country and to use these as a way of predicting which areas need the flu vaccine first. In collaboration with the CDC, they developed a website where they can select the trends by state, city, and region to find the next outbreak. The project is dead in 2014 after researchers found out that Google is overestimating the outbreak by over 50 percent. Out of all the 45 search terms, Google is using, they also found out that terms like “High School Basketball” are good predictors of the flu which is seasonal just like the flu. [6]

Target

Target has created a model that could find which shopper will be more likely to become pregnant just by looking at their shopping history and habits. People with newborns are so tired if they need something at a store, they will end up buying everything else they need. By analyzing the data, Target was able to identify 25 different items that when analyzed together would allow them to predict if someone was pregnant. Target’s program was so accurate, it could assign any regular shopper a pregnancy prediction score. The problem with this model is that target could not let on how much they knew and all shoppers got annoyed if they receive any kind of advertisement making it clear that Target is studying their reproductive plans. To partially resolve this issue, target started adding non-pregnancy related items to their email recommendations with random baby products.[7]

Elections (Mitt Romney)

In the year 2012, a Republican Presidential candidate Mitt Romney used the power of big data to analyze his chances of sitting in the White House. His team has designed a platform called ORCA for finding insights about the situation on the polling stations, which later helps their volunteers to vote. According to ABC news, the platform has crashed multiple times because the servers are overloaded because it is untested before the elections. Relying completely on the data has made Mitt Romney’s campaign team incompetent. [8]

References

1. “What's The Biggest Big Data Challenge in 2018?” Welcome to Intersog – Your App Development Partner in Chicago, accessed on 24 Nov. 2018 intersog.com/blog/whats-the-biggest-big-data-challenge-in-2018/.
2. “Big Data: A Tool for Development or a Threat to Privacy?” Privacy International accessed on 24 Nov. 2018 privacyinternational.org/blog/1434/big-data-tool-development-or-threat-privacy.
3. “Bill Schmarzo's Top Big Data, Data Science and IOT Blogs.” Cloud News and Thought Leadership, accessed on 24 Nov. 2018 cloudtweaks.com/2015/03/how-much-data-is-produced-every-day/.
4. “What Is Real-Time Analytics?” Sisense, accessed on 24 Nov. 2018 www.sisense.com/glossary/real-time-analytics/.
5. Gao, Jerry, et al. “Big Data Validation and Quality Assurance -- Issuses, Challenges, and Needs.” 2016 IEEE Symposium on Service-Oriented System Engineering, accessed on 24 Nov. 2018 (SOSE), 2016, doi:10.1109/sose.2016.63.
6. Nijhuis, Michelle. “How to Call B.S. on Big Data: A Practical Guide.” The New Yorker, The New Yorker, 19 June 2017, accessed on 25 Nov. 2018 www.newyorker.com/tech/annals-of-technology/how-to-call-bullshit-on-big-data-a-practical-guide.
7. Hill, Kashmir. “How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did.” Forbes, Forbes Magazine, 31 Mar. 2016, accessed on 25 Nov. 2018 www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/#50ccfd286668.
8. Thiessen, Marc A. “Marc Thiessen: How Obama Trumped Romney with Big Data.” The Washington Post, WP Company, 12 Nov. 2012, accessed on 25 Nov. 2018 www.washingtonpost.com/opinions/marc-thiessen-how-obama-trumped-romney-with-big-data/2012/11/12/6fa599da-2cd4-11e2-89d4-040c9330702a_story.html?noredirect=on&utm_term=.b6100406e851.