Flight arrival and departure details for all commercial flights within the USA

Mzoxolo Mbini
Dept. Computer Science
University of Pretoria
Pretoria, South Africa
u16350244@tuks.co.za

***Abstract*—Big data can be defined as a collection of data that meets a set of criteria and requires systematic, unique processing to extract relevant insights. A public dataset of commercial airline on-time performance was analyzed using the V features of big data. Volume, velocity, variety, veracity, validity, volatility, vulnerability, value, viscosity, virality, and visualization are the big V qualities used to examine the dataset. This dataset meets the requirements for volume and diversity of features for big data, and it is publicly available 30 days after the end of each month. The datasets' validity and low volatility are reinforced by their public publication and regulated collection, which only require minimum vulnerability prevention. On the other hand, the delay in data dissemination limits the ability of reactive insights to provide actionable insights**

***Index Terms*—Big Data, MapReduce, Hadoop**

I. Introduction

With advancements in technology and networking technologies, the sheer volume and variety of data available in the private and public sectors have exploded. Today's processing power enables the use of current mathematical methods to derive actionable insights from data, as the value of data has grown along with the volume and diversity of data. Industry-specific and type of insight determined the value of big data insights, but the potential value hidden in undiscovered raw data has piqued the curiosity of parties, businesses, and governments alike. Since big data is not simply a matter of size, its definition has been controversial for a long time. Doug Laney, to formalize methodologies for handling big data challenges, defined big data in 2001 as data comprising greater diversity arriving in growing volumes with ever-increasing velocities [7]. As a way of describing the various dimensions of big data, Laney's three V characteristics of big data were fashioned. Big data is characterized by its volume, variety, and velocity using Laney's definition. "Big data" has come to encompass a wide range of features since 2001. A variety of V features will be examined for a public, real-world airline on-time performance dataset compiled by the Bureau of Transportation Statistics over a two-year period, which includes all commercial flights within the United States. The V qualities assessed for the dataset in question are volume, velocity, variety, truthfulness, validity, volatility, vulnerability, value, viscosity, virality, and visualization. To find valuable feature correlations, the project's goal is to process the airline on-time performance dataset, reduce it using Hadoop map-reduce, and visualize the resultant output. Approximately 120 million records date back to October 1987, using 1.6 GB compressed storage space and 12 GB uncompressed storage space in the underlying on-time performance dataset. To optimize distribution and mapping efficiency, only a subset of data for 2007 and 2008 will be considered for future phases of the project.

Identify applicable funding agency here. If none, delete this.

II. Data exploration

An airline on-time performance dataset, the depiction of which is maintained here, was examined in terms of eleven V characteristics used to define big data. Volume, velocity, variety, veracity, validity, volatility, vulnerability, value, viscosity, virality, and visualization are among the V characteristics explored. During the exploration of the dataset, the objective of the analysis is to determine how the dataset should be processed and reduced to provide a reasonable output, from which the outputs and important relationships can be highlighted through visualisation.

A. Dataset origins

The IEEEtran class file is used to format your paper and style the text. All margins, column widths, line spaces, and text fonts are prescribed; please do not alter them. You may note peculiarities. For example, the head margin measures proportionately more than is customary. This measurement and others are deliberate, using specifications that anticipate your paper as one part of the entire proceedings, and not as an independent document. Please do not revise any of the current designations.

III. Data description with V's

A. Volume

In big data, volume refers to the amount of space a dataset occupies in addition to the number of samples and features it contains. There is no doubt that the volume of big data does not remain constant; the dataset grows according to the pace of the incoming data stream. The storage of large amounts of data can be costly due to the need for adequate infrastructure. A low-quality storage environment can affect the volatility of data [9]. With 120 million samples spanning

the years 1987 to 2008, the on-time performance dataset is made up of 1.6 GB compressed data and 12 GB uncompressed data. To increase the processing and mapping times, a subset of data accumulated between 2007 and 2008 will be used as a basis for future data processing and mapping. There were 7 453 215 samples collected in 2007, totalling 118 MB compressed data and 686 MB uncompressed data. In 2008, there were 7 009 728 samples compiled, consisting of 111 MB compressed data and 673 MB uncompressed data [3]. In total, 14 462 943 samples are present in the subgroup, requiring 1.29 GB of uncompressed physical storage space. Airports and airlines are responsible for their storage infrastructure when they produce data until the end of the month when the data is transferred to BTS. The DOT requires a particular level of security and quality within a predetermined data format, which is an indirect requirement for a quality storage infrastructure. Whenever a major airliner's plane lands or departs at a U.S. airport, data on on-time performance is accumulated over a month. Current statistics indicate that the storage infrastructure should be able to hold at least 700MB of data per month. Infrastructure design decisions must also consider variations in data velocity. As the population grows, so does the demand for travel by air, resulting in more flights departing from larger airports as a result, more data will be available.

B. Velocity

Creating, recording, transferring, and storing data at a rapid pace is known as data velocity in big data. Data can also flow continuously, necessitating the use of special equipment [5]. Speed affects the total volume of data directly. The velocity of data affects the speed of the feedback loop that controls decision-making. A dynamic data flow is crucial if a quick response is needed to the problem at hand.

A dataset's velocity is determined by the frequency with which data is delivered for that dataset. Data owner BTS does not interact directly with the technology or sensors used to gather and store data. DOT only demands that data be captured and delivered with a specific level of quality and structure per rule [8]. The practise of collecting and storing on-time performance data began in October 1987 and continues to this day. In contrast, BTS began collecting and aggregating data only in June 2003. In the United States, the main air carriers as well as all listed airports provide batch data at the end of each month for BTS to provide the pre-requisite data for flight arrivals and departures. Following the end of the month, the data are combined to produce the Air Travel Consumer Report. An entry captured on the 1st of the previous month can be 61 days old at the time of publication if it is published on the first of the subsequent month. The data is provided every month, but the carriers and airports collect it daily for each aircraft landing and departing, producing a lot of samples.

C. Variety

A dataset might be a collection of data from multiple sources, such as web pages, web log files, e-mails, sensor device data, or raw format [5]. These data might be structured, semi-structured, or raw, depending on the source. The variety measure also takes into account data with different types of features [10]. Although the data for the on-time performance dataset come from a variety of sources, it is already structured in a way required by BTS, which must be followed by all major airlines and U.S. airports. Data from BTS is compressed into comma-delimited files, such as CSV.bz2, at the time of publication.

D. Value

Researchers and businesses are attracted to the value attribute. It is pointless to collect and analyze large amounts of data if the owner does not gain actionable information from it. It is still complicated and expensive to extract, store, and process data. Consequently, insights derived from a piece of data must be greater than the costs related to task-related infrastructure [9]. Infrastructure costs include storage server expenses, processing system costs, and system maintenance charges. Big data value is determined by the intended consumer of the insight gained, as well as by company norms and laws governing permitted data risks. Data governance may not fully utilize data, either at the point of collection or at the point of action, to protect the firm against risk. To maximize potential value, big data processing creates actionable knowledge. Despite the assumption that a dataset can yield a variety of insights, each with its associated value, certain datasets do not yield deliverable insights or the insights derived have no inherent value. The purpose of gathering or developing a dataset is sometimes to confirm, verify, or expand a specific problem question. This may result in disproving, verifying, or expanding the original query. The on-time performance dataset is being processed exploratorily, without any specific problem questions in mind. For exploratory dataset processing, there is no potential value in the process. There can still be a value associated with the resulting outcomes, regardless of whether or not there is a predetermined potential value associated with the planned processing. The on-time performance dataset may be used to investigate potential problem topics outlined by the American Statistical Association [3]:

- According to the season or the day of the week, and taking into account specific airports of arrival and departure, what is the best time to take off to avoid delays?
- Is there a correlation between older plane models and flight delays?
- What are the changes in customer flight preferences over time?
- Is it possible to detect cascading failures that cause additional delays at airports?
- To lessen the impact of cascading failures, can important airports or routes be identified as vital links?

In addition to being able to use the dataset for security purposes, it can also help detect unusual flight duration patterns linked to possible terrorist activity or problems with aircraft navigation [3]. Since the set's velocity is high, taking action on generated insight will take a long time, thereby limiting the potential to derive proactive value. Moreover, the dataset

can be linked to other data structures to extract additional field insights, such as whether the weather in a certain area would affect airline delays in that area. The characteristics incorporated into the on-time performance dataset, as shown in Table I, suggest that the value derived from narrow information should be limited, as many of the features are closely related based on their descriptions. Thus, it is safe to assume that insight with inherent value will need to integrate existing variables in novel ways to provide previously uncovered data. It is consistent with the problem questions previously noted by the American Statistical Association.

E. Virality

In a peer-to-peer context, virality as a big data characteristic can provide a measure of the amount of information that is disseminated. Both the rate of dispersion and the pace of data dispersion depend on time. There is a difference between virality and big data velocity in the consumption point and the lifecycle of the data. STK releases the on-time performance dataset 30 days after the end of the previous month [8]. Thus, following the first 30-day delay, the speed at which the dataset is dispersed is determined by the dataset consumer. STK created the prior representation of the dataset's virality. When the viewpoint of dataset virality is placed on an aeroplane or an airport, the dataset disperses to a single target, the STK. According to the alternative viewpoint, dispersion occurs in the days following month-end through a data handover process to STK.

F. Veracity

By the veracity characteristic of big data, we can determine whether the data is correct. The source of data is an important determinant of its veracity, as unstructured data is intentionally extracted from sources like social media platforms, whose accuracy cannot be guaranteed. Accordingly, the value of data insights obtained concerning the problem area is directly proportional to the quality of the data. Therefore, big data is sometimes referred to by its veracity characteristic. Before the definition of this particular big data characteristic, data was considered to be clean and accurate. BTS collected on-time performance data from U.S.-based airports and major U.S.-based airlines. In addition, the data collected by BTS meets the regulatory requirements of the DOT [8]. The data veracity is therefore maintained not only by DOT requirements but also through the reports created and the data itself being made public. The accuracy of data is vital for firms that deal directly with the public or are publicly traded, such as the majority of major airlines, since it affects public perception and, as a result, share prices. As the number of data sources grows, data's authenticity is likely to dwindle as more errors are introduced [9]. Data-related errors are predicted to be limited, however, since the incoming data is actively vetted by a governmental entity, even if it is obtained from multiple agencies.

G. Validity

As with veracity characterisation, the validity of big data is dependent on the data's correctness concerning a specified application goal. However reliable and error-free the data may be, it may still be ineffective for a certain application. Correlations between features determine the quality of data analysis. Given that the current aim of the on-time performance data set is exploration, processing, and reduction without any preset problem to address, the data can be considered legitimate if it meets the veracity condition.

H. Viscosity

Large data viscosity is a measure of how difficult it is to deal with the underlying data; it is also referred to as data flow resistance [6]. In most cases, data flow resistance is connected to data collection and the complexity of data processing steps. The DOT has mandated the on-time performance dataset, so data providers should expect little to no data restrictions or opposition. In addition, the data are saved in comma-delimited files, which makes data input into various platforms much easier. To gain exploratory insights from visual and text data, more advanced processing approaches and longer processing periods are needed. However, the on-time performance data set is simple, so processing and extracting insight from it should be rather easy as long as relevant exploratory questions are formulated and addressed.

I. Visualisation

The visualization of big data occurs after the processing process is completed; however, the ability to visualize data clearly and intuitively is crucial to extracting the most value. In its unprocessed state, the on-time performance dataset has no preconfigured visualization, and from the consumer's perspective, visualization will only be worthwhile after processing has been completed.

IV. Conclusion

Data exploration has grown in popularity as a result of increasing data availability and dramatic improvements in processing power. Data in large quantities, on the other hand, has little value or utility in and of itself. Therefore, huge data can only be defined as such if it meets Doug Laney's numerous V criteria [7].

Declaration

"The University of Pretoria commits itself to producing academic work of integrity. I affirm that I am aware of and have read the Rules and Policies of the University, more specifically the Disciplinary Procedure and the Tests and Examinations Rules, which prohibit any unethical, dishonest or improper conduct during tests, assignments, examinations and/or any other forms of assessment. I am aware that no student or any other person may assist or attempt to assist another student, or obtain help, or attempt to obtain help from another student or any other person during tests, assessments, assignments, examinations and/or any other forms of assessment."

REFERENCES

[1] Punam Bedi, Vinita Jindal, and Anjali Gautam. Beginning with big data simplified. In *2014 International Conference on Data Mining and Intelligent Computing (ICDMIC)*, pages 1–7. IEEE, 2014.

[2] Jennifer Bresnick. Understanding the many v's of healthcare big data analytics. *HealthIT Analytics, https://healthitanalytics.com/news/understanding-themany-vs-of-healthcare-big-data-analytics*, 2017.

[3] Felicity L Brown, Anne M de Graaff, Jeannie Annan, and Theresa S Betancourt. Annual research review: Breaking cycles of violence–a systematic review and common practice elements analysis of psychosocial interventions for children and youth affected by armed conflict. *Journal of child psychology and psychiatry*, 58(4):507–524, 2017.

[4] ASA Data Expo. Airline on-time performance, asa section on: Statistical computing statistical graphics, 2009.

[5] Avita Katal, Mohammad Wazid, and Rayan H Goudar. Big data: issues, challenges, tools and good practices. In *2013 Sixth international conference on contemporary computing (IC3)*, pages 404–409. IEEE, 2013.

[6] Nawsher Khan, Arshi Naim, Mohammad Rashid Hussain, Quadri Noorulhasan Naveed, Naim Ahmad, and Shamimul Qamar. The 51 v's of big data: survey, technologies, characteristics, opportunities, issues and challenges. In *Proceedings of the international conference on omni-layer intelligent systems*, pages 19–24, 2019.

[7] Doug Laney et al. 3d data management: Controlling data volume, velocity and variety. *META group research note*, 6(70):1, 2001.

[8] Yoshinori Suzuki. The relationship between on-time performance and airline market share: a new approach. *Transportation Research Part E: Logistics and Transportation Review*, 36(2):139–154, 2000.

[9] Muhammad Fahim Uddin, Navarun Gupta, et al. Seven v's of big data understanding big data to extract value. In *Proceedings of the 2014 zone 1 conference of the American Society for Engineering Education*, pages 1–5. IEEE, 2014.

[10] Samuel Fosso Wamba, Shahriar Akter, Andrew Edwards, Geoffrey Chopin, and Denis Gnanzou. How 'big data'can make big impact: Findings from a systematic review and a longitudinal case study. *International Journal of Production Economics*, 165:234–246, 2015.