Development of Automated Flight Data Collection System for Air Transportation Statistics

Satria Bagus Panuntun\textsuperscript{1} and Setia Pramana\textsuperscript{1,2}\textsuperscript{*}

\textsuperscript{1}Directorate of Analysis and Statistics Development, BPS Statistics Indonesia, Jakarta, 10710, Indonesia
\textsuperscript{2}Politeknik Statistika STIS, Jakarta, 13330, Indonesia

\textsuperscript{*}E-mail: setia.pramana@stis.ac.id

Abstract. Data and information regarding air transportation is very crucial for all aspects, such as economy, people mobility and tourism. Currently, the Air Transportation Statistics is based on administrative data of different institutions. It is necessary to have a new data source regarding air transportation activities that can be used as an alternative reference for air transportation data which is faster and more granular. This research aims to study a new approach to produce air transportation statistics, especially air transportation statistics in Indonesia from big data that can be used as comparative or as complementary data for official air transportation statistics. This research using a website scraping method from a site that provides monitoring and tracking services for all flights in the world. The flight data was collected daily from the 15 busiest airports in Indonesia for both departure and arrival flights using the API provided by the website. The Scrapy module in the Python programming language is implemented. The data was collected daily from March 15, 2020, to August 31, 2020. The results of the flight data set contain information about the flight code, aircraft code, airline name, departure airport, departure city, arrival airport, arrival city, date/time of departure and arrival, and flight status. The result shows that it is feasible to use big data as a comparative or as a complementary of official statistics, especially in air transportation statistics. By using the web scraping technique, the indicator that usually requires more time and cost can be done in real-time and less cost. This new approach is expected to improve the quality of official air transportation statistics.

1. Introduction
Indonesia as the largest archipelago with a great population is faced with a big challenge in the transportation sector. The demand for air transportation services keeps increasing each year in line with the increase in population and welfare of its people. Air transportation has a dominant role, especially involved with the demand for fast transportation, which means air transportation is the only choice. The importance of the air transportation sector must be accompanied by a directed development policy in this sector so that the condition of air transportation will get better in the future. Data and information regarding air transportation activities can be used by the government to determine policies for the development of air transportation.

One of official air transportation data is produced by BPS statistics Indonesia and often called Air Transportation Statistics. It provides an overview of air transport activities between airports nationally and internationally in a certain period of the year. There are two types of data presented in the air transportation statistics produced by BPS, namely Air Transport Traffic
Data and Aircraft Registration Data. The two data presented were collected in several ways. Air Transport Traffic Data in the form of origin of destination, movement of aircraft, passengers, baggage, goods, and post/parcel is collected from airports both managed by PT. Angkasa Pura and the Directorate General of Civil Aviation Indonesia. This data is a recapitulation of the flight activities of each airport every month. Meanwhile, Aircraft Registration Data is collected from the Indonesia Civil Aviation Directorate and scheduled airline production data is obtained from the airlines [1,2].

Data and information regarding Indonesia air transportation activities produced by BPS are published through the publication of Air Transportation Statistics. This publication is released annually for data in the previous year period. This causes a delay in the delivery of information of about one year, even though data and information on air transportation are also used to complement statistics in other sectors, such as tourism statistics. Also, there is no comparative data for official air transportation statistics. Therefore, it is necessary to have a new data source regarding air transportation activities that can be used as an alternative reference for air transportation data which is faster and more granular which can be used as comparative data for official air transportation statistics.

An alternative that can be used is data provided publicly by flight tracking websites such as FlightRadar24 and FlightStatus. These sites provide information about air transportation activities more broadly and up-to-date. Information about flight schedules on the site is presented in real-time at any time for all types of flights at all airports in the world. If this information can be collected, then the results can be used as a reference source of air transportation data which is faster and cheaper.

In this study, the development of a data collection system for air transportation activities was carried out by collecting data on daily flight schedules at Indonesian airports for domestic and international flights based on sources from big data using the web scraping method. Several studies have been carried out regarding data collection using the web scraping method as a comparison or as complementary data for official statistical data. 'The potential of big data form official statistics in Indonesia by utilizing several data sources and for different purposes' [3], for example, aims to determine the use of big data (web scraping) to produce tourism statistics, especially accommodation statistics in Indonesia using online travel agent data. Also, there is 'Development of Online Travel Web Scraping for Tourism Statistics in Indonesia' [6] which uses web scraping methods from several online travel agency sites in Indonesia to obtain occupancy rates rooms and weighted average daily and monthly accommodation prices as comparative and complementary of official tourism statistics. The use of air transportation activities data from big data has also been echoed by BPS in observing the conditions of mobility that occurred during the COVID-19 pandemic, which aims to monitor the relationship between the number of flights and the current pandemic conditions [4]. In its development, the analysis of flight data in 'Study Big Data as Complete Social Statistic Data and Information' [5], will be linked to indicators such as population density and mobility.

The collection of flight data, which is sourced from big data, in this study was carried out using the web scraping method. 'Using Web Scraping In a Knowledge Environment To Build Ontologies Using Python And Scrapy' [7], uses web scraping techniques using Python and Scrapy to extract content on the e-commerce website page field to build ontologies. In this study, the web scraping method is used to retrieve many unstructured data available from the FlightStatus site, then convert it into structured data that can be processed using existing statistical methods. With the web scraping method, data collection can be done in real-time and can save time and money. The results of this data collection will later be used as comparative or as complementary data for official air transportation statistics. This research aims to study a new approach to produce air transportation statistics, especially air transportation statistics in Indonesia from big data that can be used as comparative or as complementary data for official air transportation.
statistics.

2. Methodology

2.1. Data Collection Method
In this study, the data collection was carried out by taking flight data in Indonesia through the FlightStatus By Cirium website. We chose Flightstatus By Cirium or better known as FlightStatus as the data source because this site is one of the biggest sites that provides monitoring and tracking services for all flights in the world in real-time and free of charge.

This site collects timetable, airlines, airports, and position data from more than 600 sources to generate flight status information services. The workflow of the data collection process from the FlightStatus site can be explained in 5 steps, Preparation, Determining Web Scraping Technology, Flight Data Collection, Daily Scraping, Data Cleaning and Compilation.

2.1.1. Preparation
The first step before collecting flight data is to determine the list of airports to be used as observation units using the IATA airport code (the International Air Transport Association location identifier). IATA airport codes are regulated by the International Air Carriage Association and published three times in the IATA Airline Coding Directory. This three-letter code is used to designate airports around the world. These IATA airport codes will later be used to identify airports for which data will be taken on the FlightStatus site [11].

In this study, the airports used as observation units are the 15 busiest airports in Indonesia [10]. Some of the IATA names and codes for these airports are Soekarno-Hatta International Airport, Tangerang (CGK), Ngurah Rai International Airport, Denpasar (DPS), Juanda International Airport, Surabaya (SUB), and others.

2.1.2. Determining Web Scraping Technology
After the list of IATA airport codes has been determined, the next step is to determine which web scraping technology to use. To obtain information from the FlightStats site, we use the site’s Application Programming Interface (API) that contains HTTP Headers and HTTP responses such as IATA airport codes, flight types, and flight dates that can be extracted.

First, to get the data, we use a web scraping technique using Hypertext Transfer Protocol (HTTP) Programming [8]. we send an HTTP request to the API server by sending some request parameters such as the IATA airport code that has been determined in the previous stage, type of flight (arrival or departure), and date of the flight. After sending the request and its parameters, the server will send a JSON that contains a list of links of all flight schedules at the airport and flight times according to the request sent. An example of the complete link structure of the FlightStatus web API links can be seen in Figure 1.

![flightstats.com](flightstats.com/v2/api-next/flight-tracker/arr/DPS/2020/07/20/6?carrierCode=&numHours=6)

**Figure 1.** Example of a complete API link structure.

The API structure in Figure 1 contains a request for an ARR (arrival) flight schedule to DPS Airport (Bali Ngurah Rai International Airport) on July 20, 2020. An example of the request result is in the form of a JSON file which contains the flight schedule links that can be seen in Figure 2. These links or URLs will lead us to a page that contains detailed information about flight schedules at the airport and the flight dates that have been requested before.
2.1.3. Flight Data Collection

After all the links for each flight schedule from the FlightStatus site are collected, we can start scraping the flight information from each URL. The data collected from each link on the list consists of flight code, aircraft code, airline name, departure airport, departure city, arrival airport, arrival city, departure and arrival date/time, and flight status as shown in Figure 3.

Figure 3. Example of the page that contains detailed information about flight schedules.

In this study, we use the Scrapy framework to do the job. We use Scrapy because it can run asynchronously, so the flight information gathering process can be done faster rather than using only the requests module in python. With the Scrapy module, any information in the image will be extracted by determining the location of the element in the HTML structure of the page [9]. For example the airplane code element (ID * 6514) as shown in the Figure 3, its location is located at div class = "TextHelper-s8bko4a-0 ekVwAR" in the HTML CSS structure of the page.

2.1.4. Daily Scraping

This data collection is carried out every day by collecting flight activity data on the previous
day, this can be done because of the Flightstatus site stores historical flight schedule data up to 3 (three) days before. This is also done to ensure that the status of the flight data taken is fixed at the final status of the flight, for example departed (the plane has flown according to its schedule) or canceled (the plane has canceled its flight according to its schedule).

2.1.5. Data Cleaning and Compilation
Flight data obtained from the results of the FlightStats web scraping process must be cleaned before being analyzed. The cleaning process includes:

- Changing the data field into the correct data type. Changing the data field is necessary so we can analyze the result in the right way.
- Removing duplication of data based on flight code, airline name, departure airport, arrival airport, and flight date.
- Deleting flight data with unknown status. On the FlightStatus site, there are many flights with the status "Unknown". Flight with the status of "Unknown" is a flight whose final status cannot be captured by the FlightStatus site, this happens when a flight schedule has not changed in its status for too long or it can also occur because of the flight schedule has been deleted by the relevant airline.

After the cleaning process is done, the cleaned data from a certain day is stored in the local database and the cloud. In a local database we store data in a MySQL formatted database, while in the cloud, we store the database using SQL in Bigquery. We use Bigquery because it can manage data using fast SQL-like queries for real-time analysis.

2.2. Data Analysis Method
In this study, the analysis of the collected flight data was carried out in several stages. This flight data analysis begins with data grouping, categorizing countries of origin and destination, and visualizing the results.

2.2.1. Data Grouping
The flight data set that has been collected is then grouped into flight departure and arrival flight data sets. These two data sets will be analyzed separately for departure and arrival flights.

2.2.2. Categorizing Countries of Origin and Destination
Each line in the flight data set contains information about the city of origin and destination of the flight. Each city information will then be grouped according to their respective countries so that later it can be carried out for analysis according to country entities.

2.2.3. Descriptive analysis
The tabulated data set is then analyzed by assessing its characteristics. These characteristics include the average value of daily flights per month or per time period, the number of flights per month or per time period, and others.

2.2.4. Visualizing the Results
Data that has been processed will be visualized to make it more representative and interactive. The tools used to visualize are a desktop application called Microsoft PowerBI.
3. Results and Findings

3.1. Data Collection Results

The result of the data collection process is a data set with 190,128 rows of arrival flight data and 188,890 rows of departure flight data from and to the 15 busiest airports in Indonesia. This data set is collected from all flight schedules from 15 March 2020 to 31 August 2020 for the 15 busiest airports in Indonesia. There were three days during the data collection period, namely March 19, April 1, and May 2, during which the flight data failed to retrieve due to technical errors. This caused the flight data on these dates to be unable to enter the data set. The 15 busiest airports in Indonesia include:

- Ahmad Yani International Airport (SRG) - Semarang
- Adisutjipto International Airport (JOG) - Yogyakarta
- Halim Perdanakusuma International Airport (HLP) - Jakarta
- Hang Nadim International Airport (BTH) - Batam
- Internasional Husein Sastranegara International Airport (BDO) - Bandung
- Juanda International Airport (SUB) - Surabaya
- Kualanamu International Airport (KNO) - Medan
- Lombok International Airport (LOP) - Lombok
- Ngurah Rai International Airport (DPS) - Denpasar
- Soekarno–Hatta International Airport (CGK) - Jakarta
- Sultan Aji Muhammad Sulaiman International Airport (BPN) - Balikpapan
- Sultan Hasanuddin International Airport (UPG) - Makassar
- Sultan Syarif Kasim II International Airport (PKU) - Pekanbaru
- Supadio International Airport (PNK) - Pontianak

From the collected dataset, it was found that some of the data included flights with an "unknown" status. A flight with an "unknown" status is a flight whose final status cannot be captured by the Flightstatus site, this happens when a flight schedule has not changed in its status for too long or it could happen because the flight schedule has been deleted by the airline. Table 1 below shows an example data set of the number of data on arrival flights to 5 airports with "arrived" status obtained per month of collection.

| Arrival Airport                        | March | April | May  | June | July | August |
|----------------------------------------|-------|-------|------|------|------|--------|
| Juanda Int. Airport                   | 3502  | 5390  | 225  | 3434 | 3752 | 7169   |
| Kualanamu Int. Airport                | 1504  | 2094  | 130  | 1320 | 1266 | 3196   |
| Ngurah Rai Int. Airport               | 2818  | 2736  | 214  | 2286 | 2656 | 3589   |
| Soekarno–Hatta Int. Airport           | 9486  | 14199 | 1591 | 9911 | 10468| 20550  |
| Sultan Hasanuddin Int. Airport        | 2669  | 4087  | 335  | 3110 | 3492 | 5684   |

Meanwhile, Table 2 below shows an example data set of the number of flight departure data from 5 airports obtained per month.
Table 2. Total Data on Departure Flights from 5 Airports in Indonesia (March - August 2020)

| Departure Airport                  | March | April | May  | June | July | August |
|------------------------------------|-------|-------|------|------|------|--------|
| Juanda Int. Airport                | 3562  | 5457  | 226  | 3392 | 3690 | 7180   |
| Kualanamu Int. Airport             | 1481  | 2198  | 108  | 1283 | 1303 | 3237   |
| Ngurah Rai Int. Airport            | 2329  | 2690  | 240  | 2198 | 2649 | 3614   |
| Soekarno–Hatta Int. Airport        | 9425  | 13918 | 1694 | 9678 | 10325| 20486  |
| Sultan Hasanuddin Int. Airport     | 2794  | 4362  | 351  | 3045 | 3656 | 5685   |

3.2. Data Visualization and Analysis Results

Data visualization was carried out using the Microsoft PowerBI application. The graph in Figure 4 shows the number of departing flights from the 15 busiest airports in Indonesia for domestic flights. In the Figure 4 below, it can be seen that the number of departures from the 15 busiest airports in Indonesia has decreased since mid-March 2020. The three airports with the largest drop in departures were Jakarta Soekarno Hatta Airport, Juanda International Airport Surabaya, and Sultan Hasanuddin Makassar Airport with their respective percentage decreases 71.4%, 61%, and 70.8% respectively until April 24, 2020. The most significant decrease in the number of departures occurred in May 2020. This is expected to occur due to the release of Regulation of the Indonesia Minister of Transportation Number 25 2020 concerning Transportation Control during the Idul Fitri 1441 Homecoming Season in the Context of Preventing the Spread of Covid-19 which is valid from 24 April to 31 May 2020. Then after this regulation ended, the number of domestic departing flights began to continue to increase from June 1, 2020 until August 31, 2020.

Figure 4. Number of Departure Flight from 15 Airports in Indonesia by Date of Flight (15 March – 31 August 2020).

Similar to the conditions for departures from the 15 busiest airports in Indonesia, the condition of arrival flights has also decreased since mid-March 2020 as can be seen in the graphic figure 5 below. The three airports with the largest drop in arrivals were also Jakarta Soekarno Hatta Airport, Juanda International Airport Surabaya, and Sultan Hasanuddin Makassar
Airport with their respective percentage decreases 72.5%, 60.9%, and 71.2% respectively until April 24, 2020. The enactment of Regulation of the Indonesia Minister of Transportation Number 25 of 2020 also has a significant impact on reducing the number of arrivals to the 15 busiest airports in Indonesia. Then there was an increase in the number of arrival flights after this regulation ended on June 1, 2020 until August 31, 2020.

Figure 5. Number of Arrival Flight to 15 Airports in Indonesia by Date of Flight (15 March – 31 August 2020).

4. Discussion
From the flight data set that has been collected based on sources from big data (web scraping from FlightStatus website), it can be said that the results represent the flight conditions of 15 airports in Indonesia. Although the current results cannot be compared with the official air transportation statistics, the existing data set can be used as a complement to it. With the flight data set that has been collected, data users can access information about the number of flights to and from 15 airports in Indonesia which can be filtered by type of flight, name of the airline, flight status, origin and destination airport, and date/time of flight more quickly (almost real-time). This information will certainly be useful both for air transport statistics and for other statistics. So it can be said that it is feasible to use big data as a comparative or as a complementary of official statistics, especially in air transportation statistics. By using the web scraping technique, the indicator that usually requires more time and cost can be done in real-time and less cost. This new approach would improve the quality of official air transportation statistics.

5. Conclusion
Flight data set that collected from big data using the web scraping method can be used as comparative and complementary data from official air transportation statistics or other official statistics. The results show that in August 2020 the condition of domestic flights in Indonesia has continued to increase since the end of Regulation of the Indonesia Minister of Transportation Number 25 of 2020. The flight conditions in August 2020 generally exceeded the flight conditions during the Large-Scale Social Restrictions period. As for international flights, the number of
flights per day is still likely to be stable until August 2020 since it decreased drastically in mid-March to early April 2020. This could happen because international flights are not affected by the enactment of Regulation of the Indonesia Minister of Transportation Number 25 of 2020.

References

[1] BPS Statistics Indonesia. (2017). Air Transportation Statistics 2017. Jakarta: Subdirectorate of Transportation Statistics.
[2] BPS Statistics Indonesia. (2018). Air Transportation Statistics 2018. Jakarta: Subdirectorate of Transportation Statistics.
[3] Pramana, S., Yuniarto, B., Kurniawan, R., Yordani, R., Lee, J., Amin, I., Satyaning, P.P.N.L.P., Riyadi, Y., Hasyyati, A.N., Indriani, R. (2017). Big Data for Government Policy: Potential Implementations of Big Data for Official Statistics in Indonesia. IWBIS, IEEE.
[4] Pramana, S., Bustaman, U., Niken Larasati, D., Adriana Sofa, W., Fitria Deli, N., Faris, M., Rahayu, Y., Lailatul Fitriyani, A., Bagus Panuntun, S., Yoga Paramartha, D., Yuniarti, Y., Satria Putri, Z. H., Nuria, N., Takdir, T., Mariyah, S., Luqman, M., Yesa Ermawan, G., Nori, N., Koswara, A., Puspita Sari, D., Noviyanti, I., Sinung Nugroho, A.R. (2020). Big Data Review of The Impact of COVID-19. Jakarta: Subdirectorat of Statistical Model Development BPS Statistics Indonesia.
[5] Pramana, S., Rahayu, Y., Lailatul Fitriyani, A., Adriana Sofa, W., Bagus Panuntun, S., Yoga Paramartha, D., Yuniarti, Y. (2020). Study Big Data as Complete Social Statistic Data and Information. Jakarta: Subdirectorat of Statistical Model Development BPS Statistics Indonesia.
[6] Adhinugroho, Y., Pratama Putra, A., Luqman, M., Yesa Ermawan, G., Takdir, T., Mariyah, S., Pramana, S. (2020). Development of Online Travel Web Scaping For Tourism Statistics in Indonesia. Information Research Journal.
[7] El Asikri, M., Krit, S., Chaib, H. (2020). Using Web Scaping In A Knowledge Environment To Build Ontologies Using Python And Scrapy. EJMCM.
[8] Saurkar, Anand V., Pathare, Kedar G., Gode, Shweta A. (2018). An Overview On Web Scaping Techniques And Tools. IJFRCSC.
[9] Chaves, A. (2020). Scrapy 2.3 documentation. (Python Software Foundation). Crummy https://docs.scrapy.org/en/latest.
[10] Wikipedia. List of the busiest airports in Indonesia. Accessed on July 18, 2020, from https://en.wikipedia.org/wiki/List_of_the_busiest_airports_in_indonesia.
[11] IATA. IATA Airline and Location Codes. Accessed on July 18, 2020, from https://www.iata.org/en/services/codes/.