Development of a Data Fusion Framework to support the Analysis of Aviation Big Data

Eugene Mangortey[*], Jerome Gilleron[*], Ghislain Dard [*], Olivia J. Pinon[†] and Dimitri N. Mavris[‡]
Georgia Institute of Technology, Atlanta, GA, 30332

The Federal Aviation Administration (FAA) is primarily responsible for the advancement, safety, and regulation of civil aviation, as well as overseeing the development of the air traffic control system in the United States. As such, it is faced with tremendous amounts of data on a daily basis. This data, which comes in high volumes, in various formats, from disparate sources and at various frequencies, is used by FAA analysts and researchers to make accurate forecasts, improve the safety and operational performance of their operations, and streamline processes. However, by its very nature, aviation Big Data presents a number of challenges to analysts: it impedes their ability to get a real-time picture of the state of the system, identify trends and operational patterns, make real-time predictions, etc. As such, the overarching objective of the present effort is to support FAA through the development of a data fusion framework to support the analysis of aviation Big Data. For the purpose of this research, three datasets were considered: System-Wide Information Management (SWIM) Flight Publication Data Service (SFDPS), Traffic Flow Management System (TFMS), and Meteorological Terminal Aviation Routine (METAR). The equivalent of one day of data was retrieved from each dataset, parsed and fused. A use case was then used to illustrate how a data fusion framework could be used by FAA analysts and researchers. The use case focused on predicting the occurrence of weather-related Ground Delay Programs (GDP) at the Newark (EWR), La Guardia (LGA), and Boston Logan (BOS) International Airports. This involved developing a prediction model using the Decision Tree Machine Learning technique. Evaluation metrics such as Matthew's Correlation Coefficient were then used to evaluate the model's performance. It is expected that a data fusion framework, once integrated within the FAA's Computing and Analytics Shared Services Integrated Environment (CASSIE) could be used by analysts and researchers alike to identify trends and patterns and develop efficient methods to ensure that the U.S. civil and general aviation remains the safest in the world.

I. Nomenclature

ARTCC	=	Air Route Traffic Control Center
ASIAS	=	Aviation Safety Information Analysis and Sharing
CASSIE	=	Computing and Analytics Shared Services Integrated Environment
CSV	=	Comma Separated Values
FAA	=	Federal Aviation Administration
FIXM	=	Flight Information Exchange Model
GDP	=	Ground Delay Program
HDFS	=	Hadoop Distributed File System
JSON	=	JavaScript Object Notation
METAR	=	Meteorological Terminal Aviation Routine Weather Report
SFDPS	=	System-Wide Information Management (SWIM) Flight Publication Data Service
TFMS	=	Traffic Flow Management System
TMI	=	Traffic Manageent Initiatives
XML	=	Extensible Markup Language

[*] Graduate Research Assistant, Aerospace Systems Design Laboratory, School of Aerospace Engineering, AIAA Student Member
[†] Senior Research Engineer, Aerospace Systems Design Laboratory, School of Aerospace Engineering, AIAA Member
[‡] Regents Professor for Advanced Systems Analysis, School of Aerospace Engineering, AIAA Fellow

II. Introduction and Motivation

The aviation industry generates data at record volumes, with most of the data generated and collected focusing on what happens in and around airplanes. This data, commonly referred to as Big Data, is characterized by the 'Four V's' [1]:

- Volume: Refers to the scale or amount of Big Data (data at rest). Large corporations typically generate, store and utilize terabytes, exabytes, petabytes and zettabytes of data
- Veracity: Refers to the accuracy of Big Data
- Velocity: Refers to how quickly streaming data (data in motion) is received and processed
- Variety: Refers to the different forms of Big Data. Big Data can be unstructured or structured. Unstructured data can be in the form of audio, video or text files, while structured data usually presents itself in databases with all features having a pretty well defined meaning

A. Big Data in aviation

Big data is leveraged by manufacturers, airlines and regulatory agencies alike to improve flight safety and fuel efficiency, increase profitability, define maintenance schedules, reduce disruptions, etc. Pratt and Whitney's new Geared Turbo Fan (GTF) engine, for example, produces 10 GB of data per second from 5000 sensors [2]. The major characteristics of the GTF engine are reduced noise and reduced fuel consumption. This is achieved by using Artificial Intelligence (AI) and Big Data to predict the demands of the engine and adjust thrust accordingly.

From an airline perspective, two key factors need to be scrutinized to ensure profitability: fuel efficiency and maintenance. Fuel is the second-highest expense for airlines and accounts for close to 17 percent of operating costs [3]. As such, airlines leverage Big Data to help them analyze and assess fuel usage on a per-trip basis [4]. Southwest Airlines, for example, collects data such as wind speed, ambient temperature, plane weight, and thrust directly from sensors embedded in its aircraft. The data is then fed into an analytics engine and fused with operational data around fuel, passenger, and cargo loads, as well as with weather data, to search for patterns in trip profitability. The airline hopes that utilizing Big Data will inform decisions such as adding or subtracting flights to routes, setting fuel loads for each aircraft, and selling additional passenger tickets. Such data can also be used by pilots during flights. In the case where turbulent conditions require the aircraft to adjust its speed and/or altitude, Big Data can be used to provide pilots with a detailed analysis of the extra fuel burn and cost associated with a specific altitude.

From a maintenance perspective, Boeing uses data from approximately 4,000 airplanes daily, as part of its Airplane Health Management (AHM) system. The data, comprised of in-flight measurements, mechanic write-ups, and shop findings, helps the company plan equipment maintenance with minimal disruptions to flights [5]. Similarly, GE Aviation analyzes data from several thousands of engines on a daily basis for early identification of sub-optimal parts to keep their engines running efficiently [6].

Airlines have also made huge strides by leveraging Big Data to explore the buying habits of customers. By analyzing more than 150 variables about each customer, including prior purchases and destinations, customers' actions can be predicted and dynamically personalized offers can be generated. Using Big Data in such fashion enabled United Airlines to increase its revenue from non-ticket sources, such as baggage fees and on-board food and services, by 15% [7].

The Federal Aviation Administration (FAA) also leverages Big Data on a daily basis. It generates, receives and utilizes a wide variety of datasets that include traffic flow data, flight data, weather data, etc. The FAA uses Big Data for real-time analysis to make accurate forecasts, streamline processes, identify operational patterns and inform the development of new concepts and methods aimed at improving the efficiency of their operations [8]. Big data is also used for research purposes to improve the already high level of air travel safety in the United States.

As illustrated, the growth of the aviation industry has been widely spearheaded by the continuous collection and analysis of the data generated by its many stakeholders. The amount of data made available is likely to continue to grow. Indeed, it is estimated that the next generation of aircraft will produce approximately eighty times the amount of data that current aircraft produce [9]. However, the nature of the data produced presents many challenges that need to be overcome for the data to become actionable and eventually be of value to the end-user.

B. Challenges associated with Big Data

Due to its very nature, the ingestion, storage, exploration, analysis and visualization of Big Data present many challenges. The volume of data that needs to be stored and managed requires scalable architectures and solutions, such as Hadoop and the Hadoop Distributed File System (HDFS) [10]. The format the data is coming in also needs to be taken into consideration. Raw, unstructured data, in particular, does not abide by a well-defined schema, making the tasks of extracting, transforming and loading the data difficult, and their automation, a significant challenge. In addition, the data received cannot be assumed to be 100% clean, accurate and trustworthy. Methods need to be in place to improve the veracity of the datasets by identifying and addressing inaccurate, incorrect or incomplete instances in a dataset. Data also comes from disparate sources. Because data silos are known to lead to inconsistencies, processes need to be developed to normalize, align, and integrate the data appropriately.

In the context of the FAA, being able to rapidly and efficiently make sense of the data is critical to their operations. However, the FAA, like other large entities, often lacks the tools and methods to enable them to do so. Data Fusion is one approach to help address this challenge. As discussed by Hall and Linas [11], Data Fusion "combines data from multiple sensors and related information to achieve more specific inferences than could be achieved by using a single, independent sensor." Hence, it is expected that by successfully fusing operational data with other types of data and analyzing results in real-time, or close to real-time, the FAA will be able to improve its capability to successfully uncover safety or security issues in the early stages before an incident or accident occurs [8]. In particular, it is expected that this process will help improve the integrity of the data, make it more widely accessible and eventually reduce the time FAA analysts spend processing and analyzing data. In general, Data Fusion is expected to facilitate the dissemination of information and improve collaboration within the FAA as well as between the FAA and its many partners and stakeholders.

C. Related Efforts

Ongoing relevant efforts enabled by Data Fusion include the Aviation Safety Information Analysis and Sharing (ASIAS) program which currently aggregates data (radar track, weather data, safety information, etc.) across multiple operators from both industry and government. The ASIAS program allows for the analysis of data at a national level to support the identification of safety trends and the assessment of the impact that changes would have on the operational environment [12]. In doing so, ASIAS has successfully led to the discovery of potential accidents or incidents before they occurred, leading to timely mitigation and prevention, and eventually, fewer accidents and casualties being recorded [13]. As part of their effort to enable an even safer Next Generation Air Transportation System, the FAA utilizes a Big Data platform, named the Computing and Analytics Shared Services Integrated Environment (CASSIE). CASSIE provides a flexible environment that fosters research, development, testing and collaboration across FAA stakeholders and organizations.

D. Research Objectives

The objectives of the research presented herein is two-fold:

1) Support the FAA through the development of a Data Fusion framework aimed at facilitating the analysis of aviation Big Data
2) Test the framework with a use case - Predicting the occurrence of weather-related Ground Delay Programs at the Newark (EWR), La Guardia (LGA) and Boston Logan (BOS) International Airports using the Decision Tree Machine Learning technique

The remainder of this paper will highlight datasets used for this research, the data acquisition and parsing processes, the use case development, its visualization and evaluation, and concluding remarks.

III. Datasets

In the context and scope of this research, three datasets were considered, and were received from the FAA's Computing Analytics and Shared Services Integrated Environment (CASSIE):

- System-Wide Information Management (SWIM) Flight Publication Data Service (SFDPS)
- Traffic Flow Management System (TFMS)
- Meteorological Terminal Aviation Routine (METAR) Weather Reports

1. System-Wide Information Management (SWIM) Flight Publication Data Service (SFDPS)

The System-Wide Information Management (SWIM) Flight Data Publication Service (SFDPS) provides stakeholders of the National Airspace System (NAS) with real-time en-route flight data for analytics, business processes, research, and other activities. It also provides Service-Oriented Architecture (SOA) message patterns containing data from the En-route Automation Modernization (ERAM) system which is transmitted through the Host Air Traffic Management (ATM) Data Distribution System (HADDS) [14]. 24 hourly SFDPS files in Extensible Markup Language (XML) [15] format from April 21, 2017 were received from the FAA and used for this research. Each SFDPS hourly file contained a variety of messages generated within that hour such as flight data, sector definitions, and general, free-form text messages to one or more NAS users. A breakdown of SFDPS messages generated between midnight and 1AM on April 21, 2017 is provided in Figure 1.

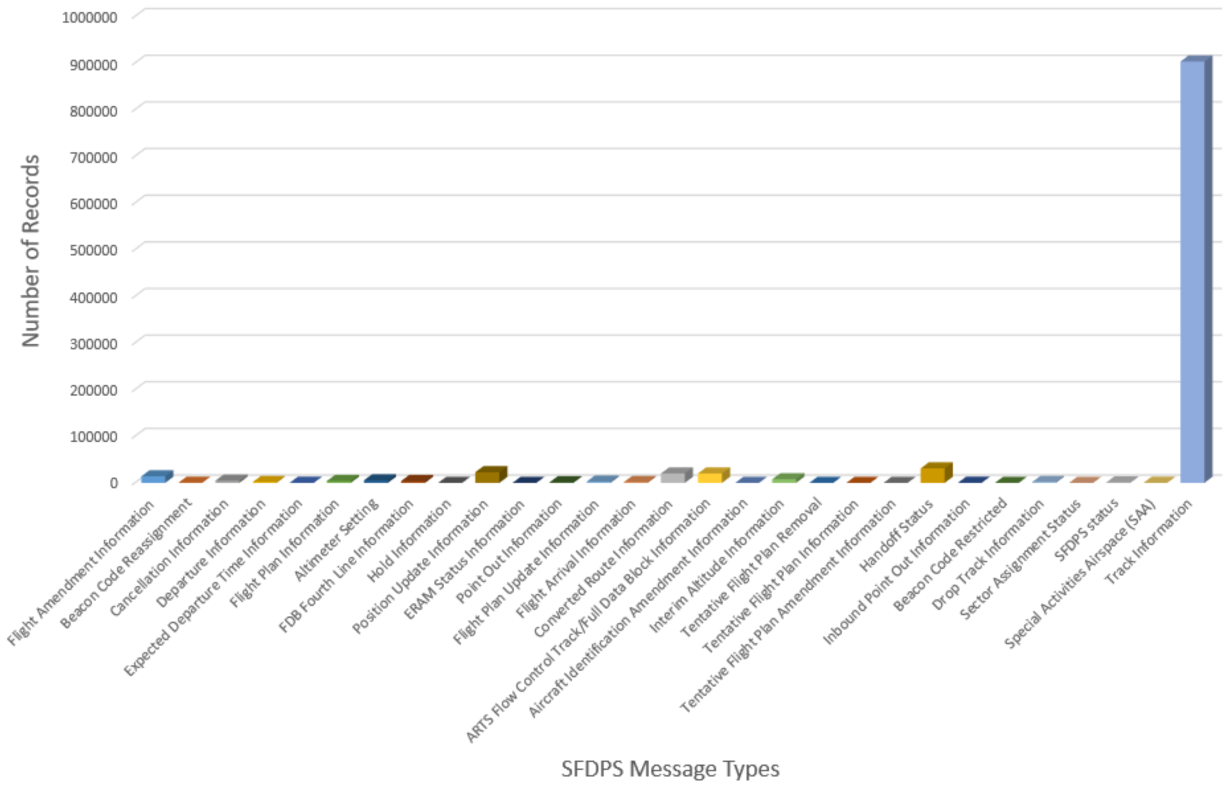


Fig. 1 Distribution of SFDPS messages from midnight to 1AM on April 21, 2017

2. Traffic Flow Management System (TFMS)

The Traffic Flow Management System (TFMS) is used by air traffic management personnel to predict traffic surges, gaps, and volume based on real-time and projected airborne aircraft [16]. TFMS is comprised of two data streams: TFMS Flow and TFMS Flight. The TFMS Flow data stream is comprised of information regarding Traffic Management Initiatives (TMI) such as Reroutes, Ground Delay Programs and Ground Stops. The TFMS Flight data stream on the other hand, provides flight track data, arrival and departure notifications, flight cancellations etc. 24 hourly TFMS Flow files from April 21, 2017 were provided by the FAA and used for the scope of this research. A breakdown of TFMS Flow messages generated between midnight and 1AM on April 21, 2017 is provided in Figure 2.

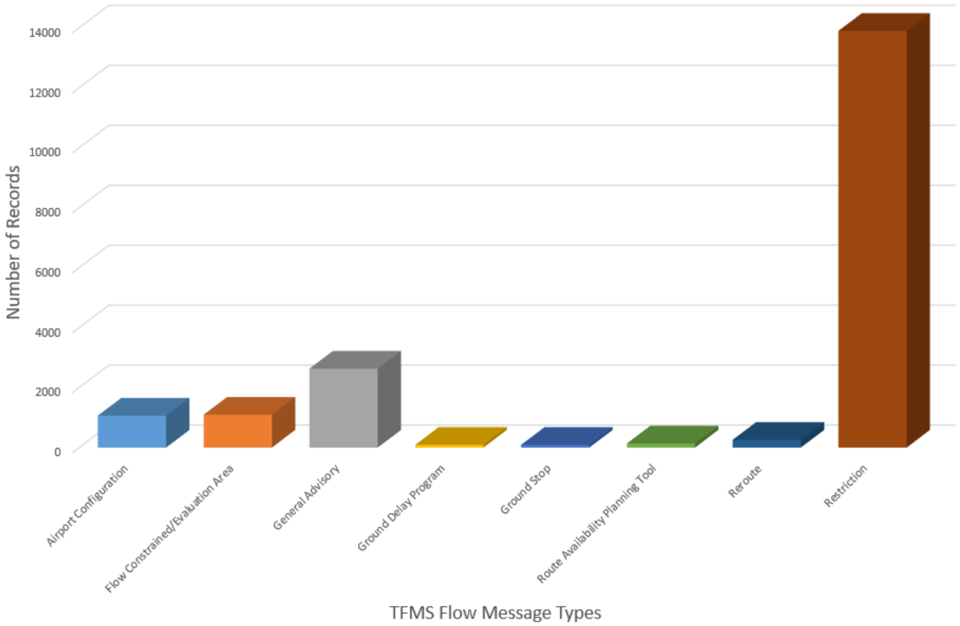
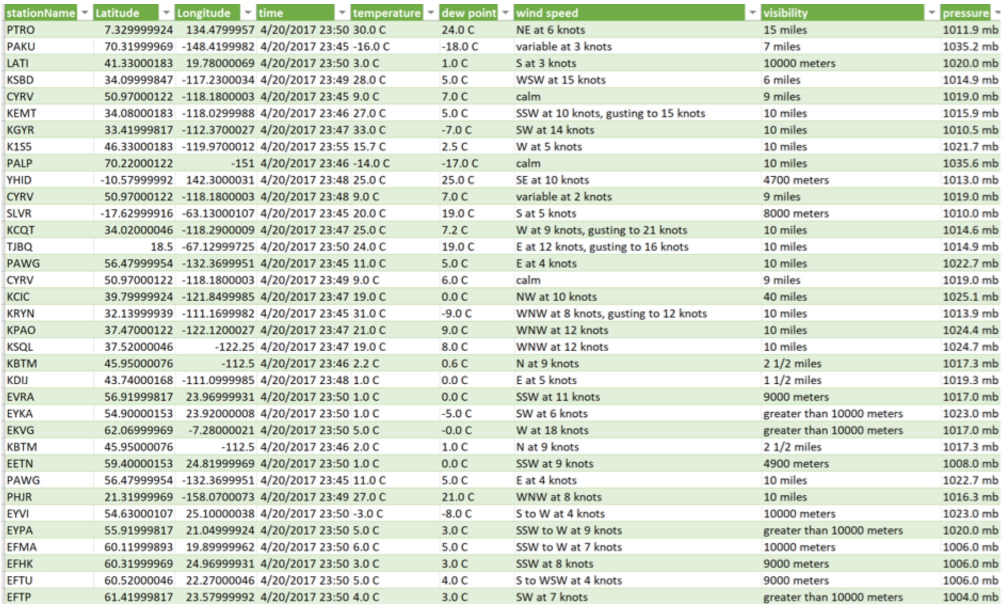


Fig. 2 Distribution of TFMS messages from midnight to 1AM on April 21, 2017

3. Meteorological Terminal Aviation Routine (METAR) Weather Reports

The Meteorological Terminal Aviation Routine (METAR) weather report is an internationally recognized format for reporting weather conditions [17]. It is often used by pilots as part of pre-flight procedures and by meteorologists for weather forecasts. These weather reports have been standardized by the International Civil Aviation Organization (ICAO) to ensure uniformity worldwide. 24 hourly METAR files were received from the FAA. Each file contained weather conditions from each hour of April 21, 2017. The METAR datasets were in NETCDF and contained relevant weather information such as wind speed, pressure, temperature, visibility etc. Figure 3 shows a decoded METAR file with METAR stations (airports) and their accompanying weather conditions.



stationName	Latitude	Longitude	time	temperature	dew point	wind speed	visibility	pressure
PTRO	7.329999924	134.4799957	4/20/2017 23:50	30.0 C	24.0 C	NE at 6 knots	15 miles	1011.9 mb
PAKU	70.31999969	-148.4199982	4/20/2017 23:45	-16.0 C	-18.0 C	variable at 3 knots	7 miles	1035.2 mb
LATI	41.33000183	19.78000069	4/20/2017 23:50	3.0 C	1.0 C	S at 3 knots	10000 meters	1020.0 mb
KSBD	34.09999847	-117.2300034	4/20/2017 23:49	28.0 C	5.0 C	WSW at 15 knots	6 miles	1014.9 mb
CYRV	50.97000122	-118.1800003	4/20/2017 23:45	9.0 C	7.0 C	calm	9 miles	1019.0 mb
KEMT	34.08000183	-118.0299988	4/20/2017 23:46	27.0 C	5.0 C	SSW at 10 knots, gusting to 15 knots	10 miles	1015.9 mb
KGYR	33.41999817	-112.3700027	4/20/2017 23:47	33.0 C	-7.0 C	SW at 14 knots	10 miles	1010.5 mb
K1S5	46.33000183	-119.9700012	4/20/2017 23:55	15.7 C	2.5 C	W at 5 knots	10 miles	1021.7 mb
PALP	70.22000122	-151	4/20/2017 23:46	-14.0 C	-17.0 C	calm	10 miles	1035.6 mb
YHID	-10.57999992	142.3000031	4/20/2017 23:48	25.0 C	25.0 C	SE at 10 knots	4700 meters	1013.0 mb
CYRV	50.97000122	-118.1800003	4/20/2017 23:48	9.0 C	7.0 C	variable at 2 knots	9 miles	1019.0 mb
SLVR	-17.62999916	-63.13000107	4/20/2017 23:45	20.0 C	19.0 C	S at 5 knots	8000 meters	1010.0 mb
KCQT	34.02000046	-118.2900009	4/20/2017 23:47	25.0 C	7.2 C	W at 9 knots, gusting to 21 knots	10 miles	1014.6 mb
TJBQ	18.5	-67.12999725	4/20/2017 23:50	24.0 C	19.0 C	E at 12 knots, gusting to 16 knots	10 miles	1014.9 mb
PAWG	56.47999954	-132.3699951	4/20/2017 23:45	11.0 C	5.0 C	E at 4 knots	10 miles	1022.7 mb
CYRV	50.97000122	-118.1800003	4/20/2017 23:49	9.0 C	6.0 C	calm	9 miles	1019.0 mb
KCIC	39.79999924	-121.8499985	4/20/2017 23:47	19.0 C	0.0 C	NW at 10 knots	40 miles	1025.1 mb
KRYN	32.13999939	-111.1699982	4/20/2017 23:45	31.0 C	-9.0 C	WNW at 8 knots, gusting to 12 knots	10 miles	1013.9 mb
KPAO	37.47000122	-122.1200027	4/20/2017 23:47	21.0 C	9.0 C	WNW at 12 knots	10 miles	1024.4 mb
KSQL	37.52000046	-122.25	4/20/2017 23:47	19.0 C	8.0 C	WNW at 12 knots	10 miles	1024.7 mb
KBTM	45.95000076	-112.5	4/20/2017 23:46	2.2 C	0.6 C	N at 9 knots	2 1/2 miles	1017.3 mb
KDIJ	43.74000168	-111.0999985	4/20/2017 23:48	1.0 C	0.0 C	E at 5 knots	1 1/2 miles	1019.3 mb
EVRA	56.91999817	23.96999931	4/20/2017 23:50	1.0 C	0.0 C	SSW at 11 knots	9000 meters	1017.0 mb
EYKA	54.90000153	23.92000008	4/20/2017 23:50	1.0 C	-5.0 C	SW at 6 knots	greater than 10000 meters	1023.0 mb
EKVG	62.06999969	-7.28000021	4/20/2017 23:50	5.0 C	-0.0 C	W at 18 knots	greater than 10000 meters	1017.0 mb
KBTM	45.95000076	-112.5	4/20/2017 23:46	2.0 C	1.0 C	N at 9 knots	2 1/2 miles	1017.3 mb
EETN	59.40000153	24.81999969	4/20/2017 23:50	1.0 C	0.0 C	SSW at 9 knots	4900 meters	1008.0 mb
PAWG	56.47999954	-132.3699951	4/20/2017 23:45	11.0 C	5.0 C	E at 4 knots	10 miles	1022.7 mb
PHJR	21.31999969	-158.0700073	4/20/2017 23:49	27.0 C	21.0 C	WNW at 8 knots	10 miles	1016.3 mb
EYVI	54.63000107	25.10000038	4/20/2017 23:50	-3.0 C	-8.0 C	S to W at 4 knots	10000 meters	1023.0 mb
EYPA	55.91999817	21.04999924	4/20/2017 23:50	5.0 C	3.0 C	SSW to W at 9 knots	greater than 10000 meters	1020.0 mb
EFMA	60.11999893	19.89999962	4/20/2017 23:50	6.0 C	5.0 C	SSW to W at 7 knots	10000 meters	1006.0 mb
EFHK	60.31999969	24.96999931	4/20/2017 23:50	3.0 C	3.0 C	SSW at 8 knots	9000 meters	1006.0 mb
EFTU	60.52000046	22.27000046	4/20/2017 23:50	5.0 C	4.0 C	S to WSW at 4 knots	9000 meters	1006.0 mb
EFTP	61.41999817	23.57999992	4/20/2017 23:50	4.0 C	3.0 C	SW at 7 knots	greater than 10000 meters	1004.0 mb

Fig. 3 A decoded METAR file with METAR stations (airports) and their accompanying weather conditions

IV. Data Acquisition and Parsing Processes

A. Data Acquisition Process

As mentioned previously, the FAA uses its Computing Analytics and Shared Services Integrated Environment (CASSIE) platform to store, process and utilize big data efficiently. CASSIE brings FAA divisions, partners and stakeholders together in an environment consisting of big data, computing power and analytical tools by using Hadoop Hortonworks for data storage and handling. Hadoop is an open-source distributed processing framework that manages data processing and storage for big data applications running in clustered systems. The main component of Hadoop used for this research is the Hadoop Distributed File System (HDFS). HDFS is a Java-based file system that provides scalable and reliable data storage and is designed to span large clusters of commodity servers with production scalability of up to 200 PB of storage. HDFS is extemely useful because a single machine has a pre-configured storage capacity. Thus, there is a limited amount of data that can be stored on the machine as the volume of data increases. The increase in data brings about the need to partition data across separate machines which is achieved by using HDFS [10].

Figure 4 highlights how data was acquired from the FAA's CASSIE platform. The System Wide Information Management System (SWIM) provides consumers with data without having to connect directly to individual data systems or sources [18]. The first step in the data acquisition process thus involved the transmission of data from the SWIM feeds. The Application Programming Interface (API), Solace Client [19], was then used to extract data from the data sources (SWIM feeds) which was then streamed to Hadoop using the open-source stream-processing software platform, Kafka [20]. NiFi processors were then used to transfer datasets in their raw format into the Hadoop Distributed File System (HDFS) for storage. Finally, the datasets were then transferred to the authors for this research.

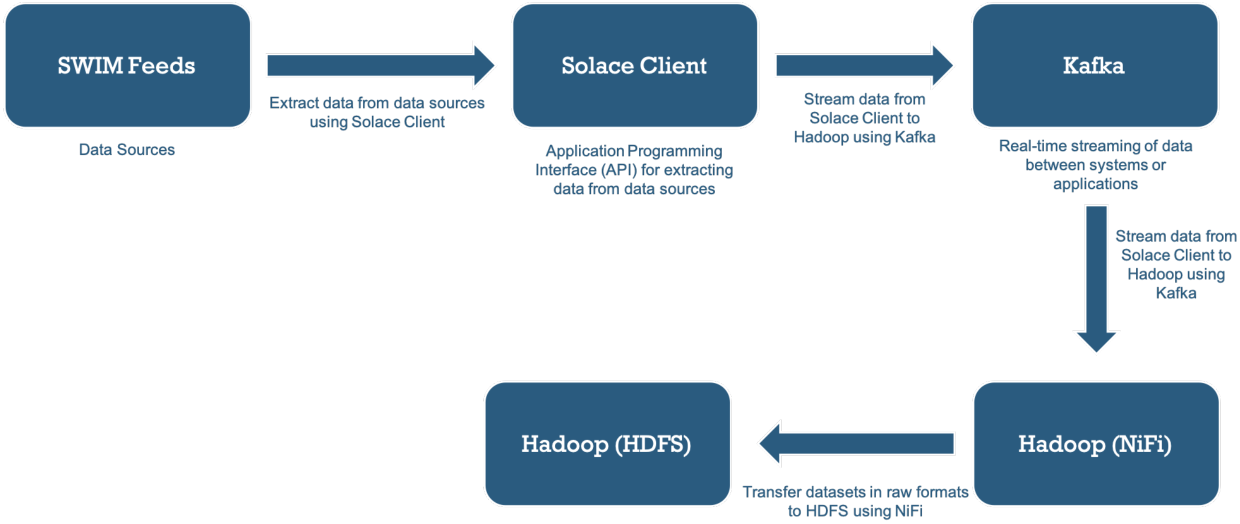

SWIM Feeds
Data Sources
Extract data from data sources using Solace Client
Solace Client
Application Programming Interface (API) for extracting data from data sources
Stream data from Solace Client to Hadoop using Kafka
Kafka
Real-time streaming of data between systems or applications
Stream data from Solace Client to Hadoop using Kafka
Hadoop (NiFi)
Transfer datasets in raw formats to HDFS using NiFi
Hadoop (HDFS)

Fig. 4 Data Acquisition Process

B. Data Parsing Process

After acquiring the datasets from the FAA, the next step involved parsing them into usable formats. The remainder of this section will focus on the steps taken to parse the datasets so as to enable data fusion across the relevant data fields.

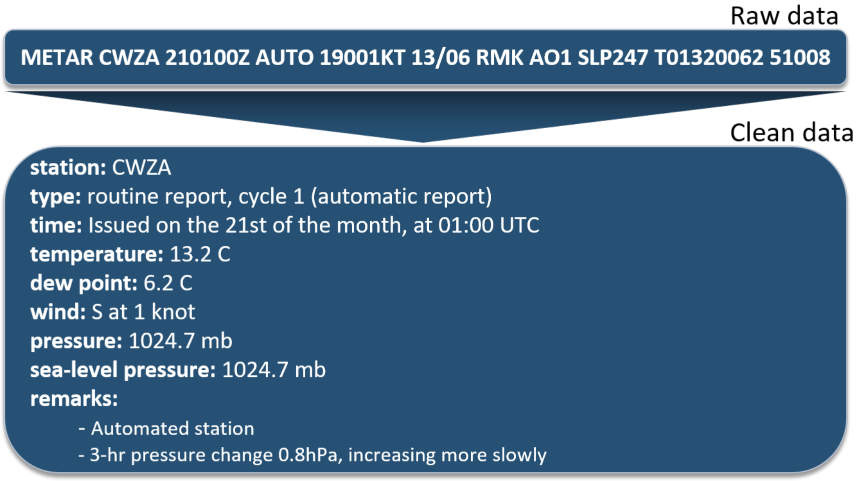
1. SFDPS and TFMS

The SFDPS and TFMS datasets were received from the FAA in the Extensible Markup Language (XML) and Flight Information Exchange Model (FIXM) formats, respectively. These two formats are much more useful for storing and sharing data than for data analytics purposes. Thus, the datasets were parsed into the JavaScript Object Notation (JSON) format which is much more usable and appropriate for data analytics purposes. The main advantage of the JSON format compared to FIXM and XML is its attribute-value structure which is similar to the structure of Python dictionaries. The SFDPS and TFMS datasets have schema files which dictate required fields and the structure of the datasets. Thus, the schema were used during the parsing process. Since the XML and FIXM formats are similar in structure, the following process was used to create Python parsers for the SFDPS and TFMS datasets:

1) Because the datasets are stored as hourly files, each file was enclosed with <root> and </root> as the header and footer of the files. This was done to establish the beginning and end of each file
2) The schema locations were then extracted from the schema files. The locations are of the format: "xlmns:......"
3) The files were then parsed using the schema and the Python module lxml ElementTree API [21]
4) The field names and their corresponding values from each message were then saved in a Python list, and the indices of the opening tags of each message were recorded
5) The Python list was then parsed using the indices of the opening tag of each message, which was then stored in a Python dictionary
6) All of the messages were appended to each other in a final dictionary which was converted into a JSON file using JSON.dumps

2. Meteorological Terminal Aviation Routine (METAR) Weather Report

The data received from the FAA was in NETCDF format [22]. Figure 5 shows the transformation of an encoded METAR message from a METAR station to the decoded message. A METAR Python module was found on GitHub [23] and was used in conjunction with a Python parser that was developed by the authors to extract and process the METAR datasets.



METAR CWZA 210100Z AUTO 19001KT 13/06 RMK AO1 SLP247 T01320062 51008

station: CWZA
type: routine report, cycle 1 (automatic report)
time: Issued on the 21st of the month, at 01:00 UTC
temperature: 13.2 C
dew point: 6.2 C
wind: S at 1 knot
pressure: 1024.7 mb
sea-level pressure: 1024.7 mb
remarks:
- Automated station
- 3-hr pressure change 0.8hPa, increasing more slowly

Fig. 5 Transformation of encoded METAR messages into decoded messages

One of the benefits of a data fusion framework is that it provides the FAA with the capability to utilize real-time and historical data for research purposes. To demonstrate this, data fusion was used in conjunction with machine learning to predict the occurrence of weather-related Ground Delay Programs (GDP) at the Newark (EWR), La Guardia (LGA) and Boston Logan (BOS) International Airports. The following section will highlight what Ground Delay Programs are, the data fusion process, the Machine Learning technique used (Decision Trees), the machine learning model generation process, evaluation metrics and the model's results.

V. Use-case Development: Predicting the occurrence of Ground Delay Programs

A. Ground Delay Programs (GDP)

Ground Delay Programs (GDP) are Traffic Management Initiatives (TMI) that are implemented to control air traffic to an airport when projected traffic demand is expected to exceed the airport's capacity for a lengthy period of time. This can be due to conditions such as inclement weather, volume constraints or runway closures [24].Whenever a Ground Delay Program is initiated, Traffic Management Personnel use the Enhanced Traffic Management System (ETMS) to predict surges and gaps in aircraft traffic based on current and anticipated airborne aircraft [25]. The Enhanced Traffic

Management System (ETMS) is then used to issue Expected Departure Clearance Times (EDCT) to affected flights. Figure 6 shows all flights affected by a weather-related Ground Delay Program at the Newark International Airport (EWR) from 00:08 GMT to 05:00 GMT on April 21, 2017 at their departure airports. Figure 7 also shows all flights affected by a weather-related Ground Delay Program at the La Guardia International Airport (LGA) from 20:08 GMT on April 20, 2017 to 02:59 GMT on April 21, 2017 at their departure airports. The flights were visualized using flight data from SFDPS and the data visualization software Tableau [26].

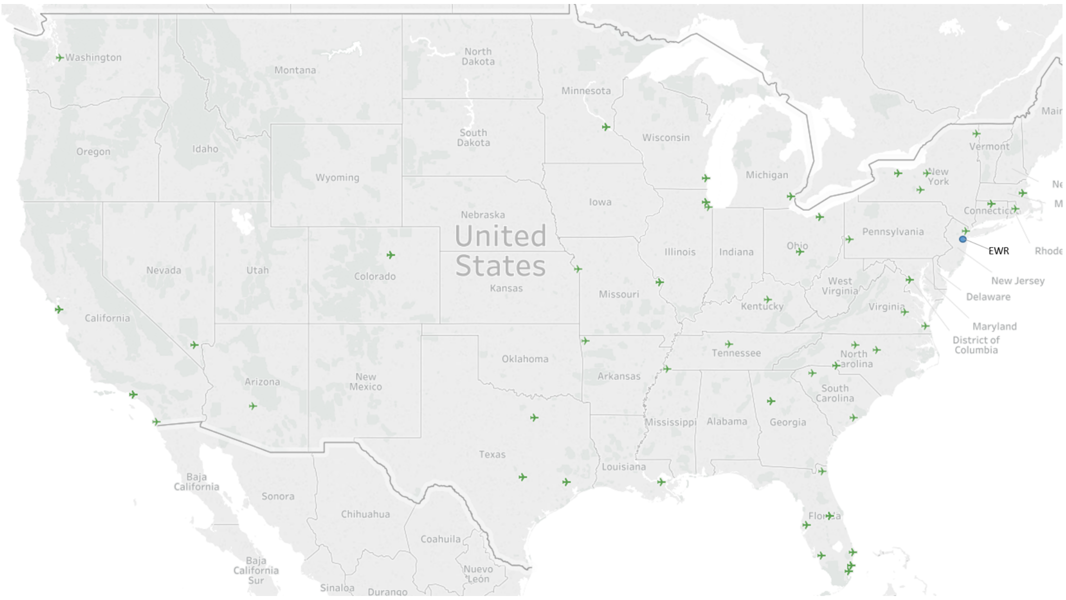


Fig. 6 Flights affected by a Ground Delay Program at Newark International Airport (EWR) from 00:08 GMT to 05:00 GMT on April 21, 2017 at their departure airports

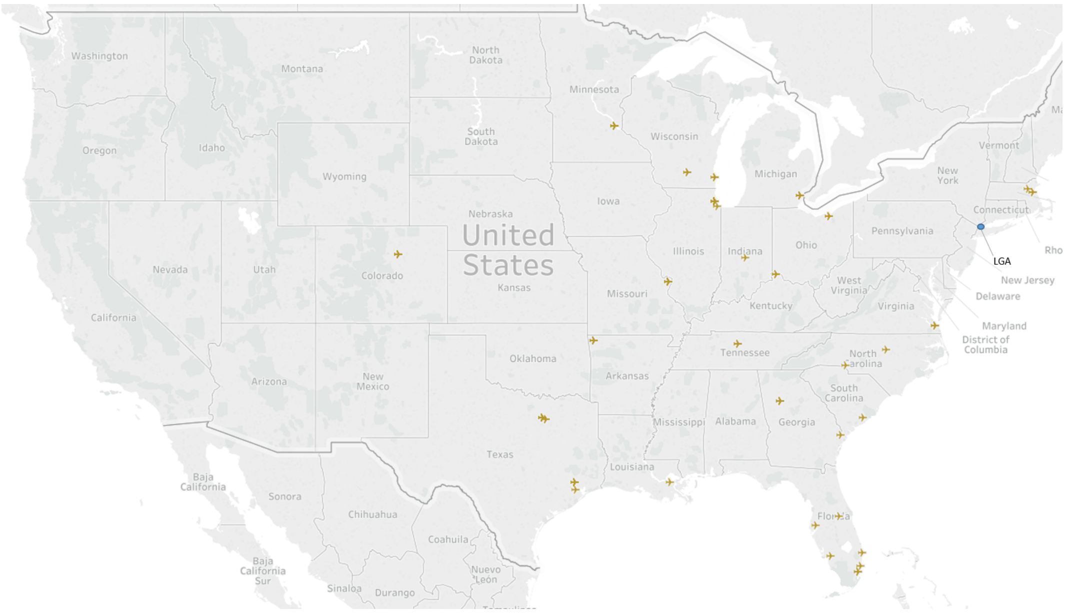


Fig. 7 Flights affected by a Ground Delay Program at La Guardia International Airport (LGA) from 20:08 GMT on April 20, 2017 to 02:59 GMT on April 21, 2017 at their departure airports

In order to successfully predict the occurrence of weather-related Ground Delay Programs (GDP), there was a need to identify common fields between the datasets and to fuse them appropriately. This was successfully achieved using Data Fusion.

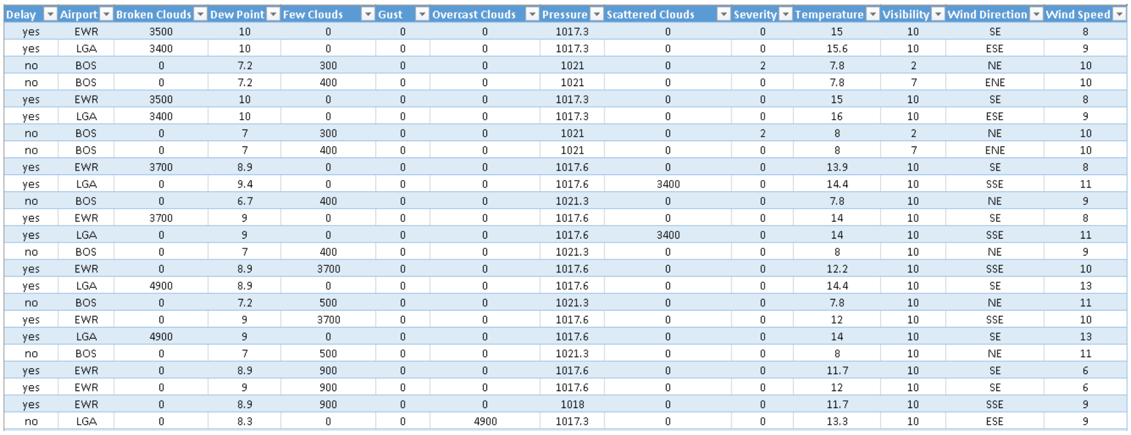
B. Data Fusion

Weather-related Ground Delay Programs, affected airports (EWR, LGA, and BOS) and their durations were extracted from TFMS. Weather conditions of the affected airports as well as the time at which they were recorded were also extracted from METAR. Thus, the datasets were fused together by airport and time. The following process was used to extract the required fields from the datasets for data fusion purposes:

1) General Advisory (GADV) messages with ’CDM GROUND DELAY PROGRAM’ in the ’advisoryTitle’ field were extracted from TFMS
2) GDP messages with ’WEATHER / THUNDERSTORMS’ in the ’IMPACTING CONDITION’ field were extracted from the General Advisory (GADV) messages
3) Active GDP messages affecting the Newark (EWR), La Guardia (LGA), and Boston Logan (BOS) International Airports were then extracted
4) Weather conditions for the affected airports were extracted from METAR
5) Data from both datasets was fused together by time and airport into a data matrix

METAR reports contain remarks on weather conditions, cloud coverage and wind speed that need to be split through a process called tokenization [27]. This involved breaking up sentences into individual words, removing punctuation, reducing words to their root form and removing stop words. After tokenization, the next step involved categorizing weather conditions by their severity levels. The categorization was done using the Aviation System Performance Metrics’ (ASPM) methodology which ranked weather conditions between 1 and 3. For the scope of this research, a severity level of 0 was assigned to weather conditions that were not recorded which meant that conditions were good. An average severity level was calculated if the remarks contained a combination of different weather conditions. The weather condition codes used in METAR reports and their corresponding severity levels as used by the Aviation System Performance Metrics (ASPM) can be found in the Appendix. Cloud coverage is also reported in METAR reports [28]. Columns for the different classifications of cloud coverage were included in the data matrix and the altitudes at which they were located were listed appropriately. The decoded METAR reports also contained a field named ’wind speed’. This field contained data on wind speed, wind direction and wind gusts at the METAR station (airport). Tokenization was used to parse this field into separate fields corresponding to wind speed, wind direction and wind gusts.

Figure 8 provides a subset of the data matrix that was created after the data fusion process.



Delay	Airport	Broken Clouds	Dew Point	Few Clouds	Gust	Overcast Clouds	Pressure	Scattered Clouds	Severity	Temperature	Visibility	Wind Direction	Wind Speed
yes	EWR	3500	10	0	0	0	1017.3	0	0	15	10	SE	8
yes	LGA	3400	10	0	0	0	1017.3	0	0	15.6	10	ESE	9
no	BOS	0	7.2	300	0	0	1021	0	2	7.8	2	NE	10
no	BOS	0	7.2	400	0	0	1021	0	0	7.8	7	ENE	10
yes	EWR	3500	10	0	0	0	1017.3	0	0	15	10	SE	8
yes	LGA	3400	10	0	0	0	1017.3	0	0	16	10	ESE	9
no	BOS	0	7	300	0	0	1021	0	2	8	2	NE	10
no	BOS	0	7	400	0	0	1021	0	0	8	7	ENE	10
yes	EWR	3700	8.9	0	0	0	1017.6	0	0	13.9	10	SE	8
yes	LGA	0	9.4	0	0	0	1017.6	3400	0	14.4	10	SSE	11
no	BOS	0	6.7	400	0	0	1021.3	0	0	7.8	10	NE	9
yes	EWR	3700	9	0	0	0	1017.6	0	0	14	10	SE	8
yes	LGA	0	9	0	0	0	1017.6	3400	0	14	10	SSE	11
no	BOS	0	7	400	0	0	1021.3	0	0	8	10	NE	9
yes	EWR	0	8.9	3700	0	0	1017.6	0	0	12.2	10	SSE	10
yes	LGA	4900	8.9	0	0	0	1017.6	0	0	14.4	10	SE	13
no	BOS	0	7.2	500	0	0	1021.3	0	0	7.8	10	NE	11
yes	EWR	0	9	3700	0	0	1017.6	0	0	12	10	SSE	10
yes	LGA	4900	9	0	0	0	1017.6	0	0	14	10	SE	13
no	BOS	0	7	500	0	0	1021.3	0	0	8	10	NE	11
yes	EWR	0	8.9	900	0	0	1017.6	0	0	11.7	10	SE	6
yes	EWR	0	9	900	0	0	1017.6	0	0	12	10	SE	6
yes	EWR	0	8.9	900	0	0	1018	0	0	11.7	10	SSE	9
no	LGA	0	8.3	0	0	4900	1017.3	0	0	13.3	10	ESE	9

Fig. 8 Data matrix obtained from data fusion process

After the Data Fusion process, the Decision Tree Machine Learning technique was used to develop a model for predicting the occurrence of weather-related Ground Delay Programs (GDP). The choice of Decision Trees in this study was justified by its main advantages: robustness to noisy data, simple and easy to understand, and it is also viewed as a 'white-box model', as results are not hidden and the process can be followed.

C. Machine Learning Technique: Decision Trees

The Decision Tree Machine Learning technique can be simply understood as a tree-like model of decisions and their potential consequences. Decision trees are intuitive tools that have been used to solve complex and strategic challenges. They have also been used to simplify problems and evaluate the cost-effectiveness of research and business decisions. A decision can be drawn (by hand or using a computer) as a flowchart-like structure. Each node represents a test on an attribute (i.e. 'if this attribute has this value X or that value Y'). This test is an internal node and represents a condition. Two or more branches or edges represent the results of one test. The end of a branch (when a series of branch ends) is a leaf or decision [29].

D. Model Generation, Validation and Testing

After successfully fusing the TFMS and METAR datasets together, the Decision Tree Machine Learning technique was applied to predict the occurrence of weather-related Ground Delay Programs using R [30]. The data matrix created after fusing the datasets was then randomly split into three sets: training, validation, and testing. Half of the data was assigned to the training set which was used to train the model, a fourth of the data was assigned to the validation set which was used to iterate and refine the model, and the last fourth of the data was assigned to the testing set to generate predictions for evaluation. The model was created using the C5.0 Decision Tree algorithm [27]. This algorithm has an 'adaptive boosting' feature which creates multiple decision trees after which the best classification of factors is selected to create the final Decision Tree. This was done by having ten trials or iterations which basically means creating ten different Decision Trees. Ten trials or iterations was used because research has showed that this reduces error rates on the test set by approximately 25% [27]. However, the algorithm stopped creating trees after the ninth iteration because adding a tenth iteration would not improve the accuracy of the model. The target of this model was whether a Ground Delay Program occurred or not and was classified as either 'yes' or 'no'. Thirteen variables including the affected airports (EWR, BOS, LGA), weather severity levels, cloud coverage altitudes and weather conditions served as predictors.

E. Evaluation Metrics

The final step involved evaluating the model's performance. Evaluating the performance of machine learning models is critical as it is important to forecast how a model will perform on data that was not used to train the model. Predicting the occurrence of weather-related Ground Delay Programs is a classification problem and was thus evaluated using metrics obtained from a confusion matrix. A confusion matrix as seen in Table 1 is a table that categorizes predictions according to whether they match the actual value.

Table 1 Confusion Matrix

	Predicted: No	**Predicted: Yes**
Actual: No	True Negative (TN)	False Positive (FP)
Actual: Yes	False Negative (FN)	True Positive (TP)

The following metrics were calculated and used to evaluate the model's performance:

1. Accuracy

This refers to the ratio of the number of true positives and negatives, and the total number of predictions [27] and is specified as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

2. Error Rate

This refers to the proportion of incorrectly classified examples [27] and is specified as:

$$Error\ Rate = \frac{FP + FN}{TP + TN + FP + FN} = 1 - Accuracy$$

3. Sensitivity

This refers to the proportion of true positives that were correctly classified [27] and is specified as:

$$Sensitivity = \frac{TP}{TP + FN}$$

4. Specificity

This refers to the proportion of negative examples that were correctly classified [27] and is specified as:

$$Specificity = \frac{TN}{FP + TN}$$

5. Precision

This refers to the proportion of positive examples that were truly positive [27] and is specified as:

$$Precision = \frac{TP}{FP + TP}$$

6. Recall

This refers to the ratio of true positives to total number of positives [27] and is specified as:

$$Recall = \frac{TP}{TP + FN}$$

7. Matthew's Correlation Coefficient (MCC)

Matthew's Correlation Coefficient has a maximum value of 1 corresponding to perfect prediction and a minimum value of -1 corresponding to total contradiction. Matthew's Correlation Coefficient is specified as:

$$MCC = \frac{(TPTN) - (FPFN)}{\sqrt[2]{(TP + FN)(TP + FP)(FP + TN)(TN + FN)}}$$

F. Results

As mentioned previously, nine boosting iterations were used in creating the model resulting in an average tree size of 7.1, which means that an average of 7.1 decision steps were taken to classify the occurrence of Ground Delay Programs. Figure 9 shows a breakdown of how the algorithm used the different predictors to predict the target value. It can be seen that the highest weighted predictors were Dew point, Wind Gusts, Altitude of Overcast Clouds, Pressure, Visibility and Wind Speed.

Table 2 provides the confusion matrix of the training set. In particular, it shows that the model accurately classified the occurrence of weather-related Ground Delay Programs at the Newark (EWR), La Guardia (LGA), and Boston Logan (BOS) International Airports.

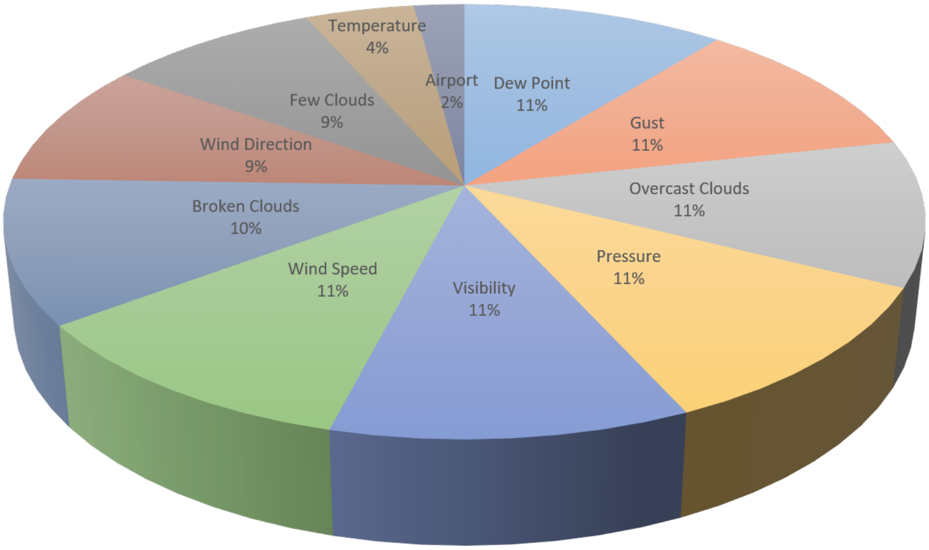


Fig. 9 Breakdown of how the algorithm used the different predictors to predict the target value

Table 2 Confusion matrix for training set

	Predicted: No	**Predicted: Yes**
Actual: No	53	0
Actual: Yes	0	67

The next step involved using the validation set to refine the model if necessary. Tables 3 and 4 summarize the confusion matrix and the results of the evaluation metrics of the validation set. The evaluation metrics in Table 4 show that the model performed very well with the validation set. Thus, there was no need to refine the model.

Table 3 Confusion matrix for the validation set

	Predicted: No	**Predicted: Yes**
Actual: No	17	2
Actual: Yes	1	40

The final step involved using the test set to evaluate the model's performance. Tables 5 and 6 provide the confusion matrix and the results of the evaluation metrics of the test set. The evaluation metrics in Table 6 show that the model performed well with the test set. The accuracy of this prediction model can be improved by increasing the amount of data used. This can be achieved once this framework is integrated in the FAA's Computing Analytics and Shared Services Integrated Environment (CASSIE).

Table 4 Evaluation metrics for validation set

Metric	Value
95% Confidence Interval	(0.8608, 0.9896)
Accuracy	0.95
Error Rate	0.05
Sensitivity	0.9756
Specificity	0.8947
Precision	0.9524
Recall	0.9756
MCC	0.88

Table 5 Confusion matrix for the test set

	Predicted: No	**Predicted: Yes**
Actual: No	17	7
Actual: Yes	1	36

Table 6 Evaluation metrics for test set

Metric	Value
95% Confidence Interval	(0.7578, 0.9416)
Accuracy	0.8689
Error Rate	0.1311
Sensitivity	0.9730
Specificity	0.7083
Precision	0.8372
Recall	0.9730
MCC	0.729

VI. Conclusion

The FAA generates, receives and uses Big Data on a daily basis. The many challenges that FAA analysts and researchers face when using Big Data may be eased using Data Fusion. As such, this research outlined steps taken by the authors to develop a Data Fusion framework by parsing, analyzing, and fusing data from three datasets: System Wide Information Management (SWIM) Flight Publication Data Service (SFDPS), Traffic Flow Management System (TFMS), and Meteorological Terminal Aviation Routine (METAR) weather reports. These datasets were received from the FAA's Computing Analytics and Shared Services Integrated Environment (CASSIE) which brings FAA divisions, partners and stakeholders together in an environment consisting of big data, computing power and analytical tools. The SFDPS dataset contained enroute flight data which was used for this research. The TFMS dataset, on the other hand, contained Traffic Management Initiatives such as Ground Delay Programs (GDP) that were issued to address constraints in the National Airspace System. The METAR dataset contained weather conditions recorded at METAR stations, i.e. airports. The SFDPS and TFMS datasets were parsed from XML and FIXM formats into JSON format which is much more appropriate for data analytics purposes using Python scripts. The METAR datasets were parsed from NetCDF format to csv format using a METAR Python package. The framework was then tested on a use case to illustrate the relevance and importance of Data Fusion in the context of predicting the occurrence of weather-related Ground Delay Programs (GDP) at the Newark (EWR), La Guardia (LGA), and Boston Logan (BOS) International airports. In order to do this, a prediction model was developed by fusing the datasets and applying the Decision Tree Machine Learning technique to the data matrix generated after the data fusion process. Confusion matrices and evaluation metrics such as Matthew's Correlation Coefficient were then used to evaluate the model's performance. It is expected that a data fusion framework, once integrated within the FAA's Computing and Analytics Shared Services Integrated Environment (CASSIE) could be used by analysts and researchers alike to identify trends and patterns and develop efficient methods to ensure that the U.S. civil and general aviation remains the safest in the world.

Acknowledgments

This research was carried out as a result of the collaboration between Georgia Institute of Technology's Aerospace Systems Design Laboratory (ASDL) and the Federal Aviation Administration's Technical Center. The authors wish to acknowledge the support of the Federal Aviation Administration's Technical Center in spearheading this research. They also wish to thank Mr. Tom Tessitore, Mr. Mike Paglione, Mr. David Chong, Mrs. Anya Berges and Mr. Rob Liguori of the FAA for helping shape this research and providing valuable feedback. The views and findings expressed in this document are those of the authors only, and do not represent those of the FAA.

References

[1] Marr, Bernard, “What Is Big Data? A Super Simple Explanation for Everyone.” , 2012. URL http://www.space.com/2995-nasa-jet-bears-nose-grows-sonic-boom-tests.html.

[2] Bhoopathi Rapolu and AviationWeek.com, “Internet Of Aircraft Things: An Industry Set To Be Transformed,” , 2016. URL http://aviationweek.com/connected-aerospace/internet-aircraft-things-industry-set-be-transformed.

[3] Airlines For America, “A4A Passenger Airline Cost Index (PACI),” , 2018. URL http://airlines.org/dataset/a4a-quarterly-passenger-airline-cost-index-u-s-passenger-airlines/.

[4] Bradbury, Danny, “How Big Data in Aviation Is Transforming the Industry,” , 2018. URL https://hortonworks.com/article/how-big-data-in-aviation-is-transforming-the-industry/.

[5] Maintenance, A., “Big Data Takes Off But Flight Is Just Beginning,” , 2017. URL https://www.avm-mag.com/big-data-takes-off-flight-just-beginning/.

[6] Tyagi, P., and Demirkan, H., “Data Lakes: The biggest big data challenges - Why data lakes are an important piece of the overall big data strategy.” , 2016. URL http://analytics-magazine.org/data-lakes-biggest-big-data-challenges/.

[7] Noyes, Katherine, “For the airline industry, big data is cleared for take-off,” , 2014. URL http://fortune.com/2014/06/19/big-data-airline-industry/.

[8] Hughes, D., “The FAA Eyes Big Data Possibilities,” , 2018. URL https://www.atca.org/Big-Data.

[9] Krishnan, V., Murray, G., and Smiley, J., “The Data Science Revolution That’s Transforming Aviation,” , 2018. URL https://www.forbes.com/sites/oliverwyman/2017/06/16/the-data-science-revolution-transforming-aviation/.

[10] Apache, “Welcome to Apache™ Hadoop!” , 2014. URL hadoop.apache.org/.

[11] Hall, David and Llinas, James, “An Introduction to Multisensor Data Fusion,” *Proceedings of the IEEE*, Vol. 85, 1997, pp. 6 – 23. doi:10.1109/5.554205.

[12] Aviation Safety Information Analysis and Sharing (ASIAS), “ASIAS Overview,” , 2018. URL asias.aero/web/guest/overview.

[13] Hueto, G., “ASIAS Overview,” , 2013. URL https://www.icao.int/APAC/Meetings/2013_APRAST3/ASIAS%20brief%20GH%2005%202013.pdf.

[14] National Transportation Systems Center, *SWIM Flight Data Publication Service (SFDPS) Reference Manual*, National Transportation Systems Center, 2015.

[15] World Wide Web Consortium, “Extensible Markup Language (XML) 1.0 (Fifth Edition),” , 2013. URL https://www.w3.org/TR/xml/.

[16] Federal Aviation Administration, “Traffic Flow Management System Reference Manual. 8.9 version,” , 2011. URL tfmlearning.faa.gov/TFMS_8.9_Reference_Manual.pdf.

[17] Wayback Machine, “Chapter 7. Safety of Flight,” , 2009. URL web.archive.org/web/20090905114733/http://www.faa.gov/air_traffic/publications/ATpubs/AIM/Chap7/aim0701.html.

[18] FAA Communications,Information & Network Programs Group (CINP), *External Consumer Administration Access to FAA Data via SWIM*, Federal Aviation Administration, 2017.

[19] HA Group Configuration, “Solace Messaging APIs.” , 2018. docs.solace.com/Solace-PubSub-Messaging-APIs/Solace-APIs-Overview.htm.

[20] Apache Kafka, “Apache Kafka,” , 2018. URL kafka.apache.org/.

[21] Python Software Foundation, “19.7. Xml.etree.ElementTree The ElementTree XML API,” , 2018. URL docs.python.org/2/library/xml.etree.elementtree.html.

[22] Unidata, “What is NetCDF?” , 2018. URL https://www.unidata.ucar.edu/software/netcdf/docs/.

[23] Pollard, T., “python-Metar,” , 2016. URL https://github.com/tomp/python-metar/blob/master/sample.py.

[24] Robyn, D., *Reforming the air traffic control system to promote efficiency and reduce delays*, The Brattle Group, 2007.

[25] Federal Aviation Administration, "Enhanced Traffic Management System," , 2017. URL https://www.fly.faa.gov/Products/Information/ETMS/etms.html.

[26] "Tableau Software," , 2018. URL https://www.tableau.com/.

[27] Lantz, Brett, *Machine Learning with R: Discover How to Build Machine Learning Algorithms, Prepare Data, and Dig Deep into Data Prediction Techniques with R*, Packt Publishing, 2015.

[28] Dr. Eric Kostelich, "METAR Format (FM-15) Surface Meterological Airways Format," , 2018. URL https://math.la.asu.edu/~eric/workshop/METAR.html.

[29] Mitchell, Tom M., *Machine Learning*, McGraw-Hill Science/Engineering/Math, 1997.

[30] The R Foundation, "The R Project for Statistical Computing," , 2018. URL https://www.r-project.org/.

[31] Office of Aviation Policy and Plans, "Weather Classification," *Federal Aviation Administration*, 2005.

Appendix

Table 7 Weather Severity Methodology Used in ASPM [31]

Weather Condition	Severity Level
Thunderstorm	3
Freezing	3
Hail	3
Snow Pellets	3
Snow	3
Snow Grains	3
Funnel Cloud	3
Fog	3
Ice Crystals	3
Ice Pellets	3
Duststorm	3
Heavy	3
Dust Whirls	3
Sandstorm	3
Mist	2
Smoke	2
Sand	2
Rain	2
Blowing	2
Widespread dust	1
Drizzle	1
Unknown Precipitation	1
Haze	1
Volcanic	1
Showers	1
Spray	1
Squall	1
Low Drifting Snow	1
Shallow Fog	1
Partial	1
Patches	1
Vicinity	1