Unidata Community Equipment Awards Proposal Cover Sheet

Proposal Title: Hardware Enhancements to Support Physics Informed Machine Learning (PIML) to Improve Airport Weather Forecasts

Date: 3/14/2023

Principal Investigator Name: Dr. Christopher G. Herbster
Title: Associate Professor of Meteorology

Co-Principal Investigator Name: Dr. Richard S. Stansbury
Title: Associate Professor of Computer Engineering and Computer Science

Institution: Embry-Riddle Aeronautical University – Daytona Beach
Telephone number: (386) 226-6444
FAX number: (386) 226-7739
Street Address: 1 Aerospace Boulevard
Daytona Beach, FL 32114

Email address: herbstec@erau.edu stansbur@erau.edu

Signature of PI: Christopher G. Herbster Digitally signed by Christopher G. Herbster Date: 2023.03.14 18:35:03 -04'00'

Name of Institution Official: Nanette Guzman
Title: Director of the Office of Sponsored Research Administration
Telephone number: (386) 226-7695
FAX number: (386) 226-7934
Email address: guzmann2@erau.edu

Signature of University Official: Nanette Guzman Digitally signed by Nanette Guzman Date: 2023.03.15 09:47:07 -04'00'

Hardware Enhancements to Support Physics Informed Machine Learning (PIML) to Improve Airport Weather Forecasts

Project Summary

The Departments of Applied Aviation Sciences and Electrical Engineering and Computer Science at Embry-Riddle Aeronautical University (ERAU), Daytona Beach campus request funds to support cluster computing for the Application of Physics Informed Machine Learning (PIML) for Weather-Induced Flight Delay Prediction project. This is an ongoing project that seeks to utilize computing resources to prepare a large-scale data set of correlated flight, airport, and weather records.

If awarded, the funds requested will enhance our existing computing infrastructure, and it will allow us to greatly increase our data processing and analysis efforts.

Project Description and Goals

The project aims to predict weather events that would disrupt aviation operations based on the history of weather impacts at the associated airports. This includes forecasting weather events that would result in the implementation of various Federal Aviation Administration (FAA) Traffic Management Initiatives (TMIs) including, but not limited to, ground delay program (GDP), Ground stop (GS), flight rerouting, airspace flow program (AFP), and deicing at the airports.

Over the years, ERAU has collected and logged an enormous amount of National Airspace System (NAS) data at the NextGen ERAU Applied Research (NEAR) Laboratory. Activity within the US NAS is logged via the Traffic Flow Management System (TFMS) messages. In compressed form, nearly 25 GB of data are logged every 24 hours which is greater than 290 GB in uncompressed XML format, with a yearly estimate of the data size reaching 106 Terabytes. Additionally, the Meteorology Program in the Department of Applied Aviation Sciences (AAS) at ERAU retrieves and archives weather data through several National Oceanic and Atmospheric Administration (NOAA) sources, as well as participation in the Unidata Internet Data Delivery (IDD) data access and sharing program. One example of storage requirements is data from the current NOAA Global Forecast System (GFS). A typical volume from one forecast cycle is more than 100 GB, four times per day, for a typical volume that exceeds 425 GB. Therefore, a month of data would exceed 12 TB and a year will exceed 155 TB.

The NEAR lab collects Traffic Flow Management System (TFMS) messages, which are used by the air traffic management systems to exchange traffic flow management data between NAS users (internal and external). The relevant TFMS messages include flight data and flow information. The flight data includes near real-time information about flight plans, flight modifications, flight tracks, departure and arrival time notifications, boundary crossings, flight plan cancellations, etc. The TFMS flow data includes information about TMIs including reroutes, ground stops, and ground delay programs [1]. The approximate data size of a year's worth of TFMS data is 12 terabytes compressed.

Currently, the HPC resources of the AAS department's Meteorology Program at ERAU are being utilized to process the data. As part of the previous efforts to utilize state-of-the-art big data solutions to process the data discussed above, tools such as Apache NiFi, Hadoop Distributed File System (HDFS), and Hive are being used to extract, transform and load required messages from TFMS. These messages include GDP, GS, Flight Plan Information, Flight Modify, Flight Plan Cancellation, DICE, APTC, Reroute, and AFP. The current computing environment is a single HP ProLiant server, with 128 GB RAM, 2x32 CPU (64 cores), and largely remote NFS shared storage. Figure 1 shows a NiFi workflow where TFMS and GFS data are being processed in multiple data pipelines.

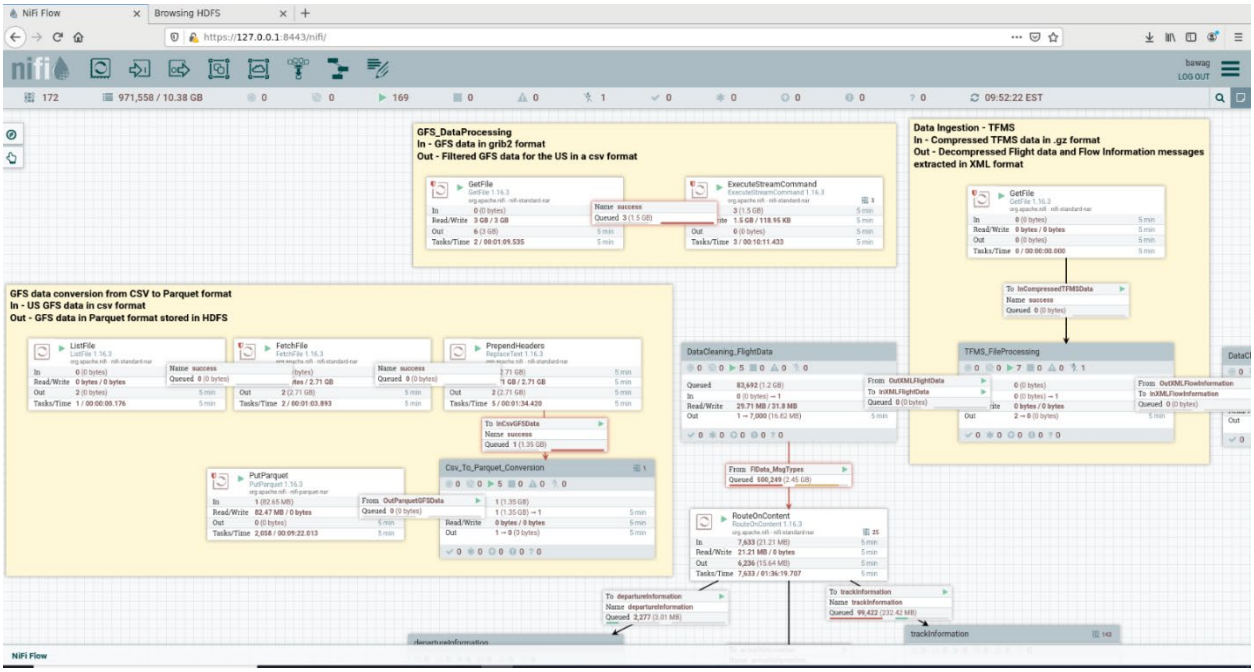

Figure 1Screenshot of Apache NiFi workflow

However, while processing such a huge dataset on a monolithic system, we have hit the processing limits of the system. Big data solutions like Apache NiFi that can automate the data flow between software systems are designed to take full advantage of the available hardware resources. This includes all the available cores, threads, RAM, and disk space. It has been observed that by setting up the big data solutions on a single server, the tools are not able to efficiently utilize large amounts of available RAM.

Under the settings shown in Figure 1, NiFi was assigned 256 threads. In the screenshot the number of active threads is shown at the top left-hand corner, here only 172 threads are active in the system, with 143 threads alone utilized by the processors working on “track information” messages from the TFMS. Due to the size of the dataset and millions of messages generated per day by the flights and airports across the airspace, the current system reached a point at which the rate of the data/number of files or messages being processed is limited, halting the upstream connections until the queues in the downstream connections fall below their performance threshold. The impact of this bottleneck is increasing the overall processing time of the messages, not providing sufficient throughput, and causing load balancing challenges. Under the current settings, it takes approximately 3-4 days to extract and process one month’s worth of TFMS data alone. Therefore, it is evident that using a single server to process such big data is not enough to process the amount of data we have.

A clustered server setup would allow big data solutions (like Apache NiFi) to take full advantage of the available resources and increased throughput. With the requested funds, ERAU shall procure and set up a cluster environment, using both new and existing computing resources, to process the flight data from TFMS, METAR weather observations at the airports, and forecast data from various NOAA forecast models. The final product will be shared with the community via a method that is still to be designed. At the least, we intend to share access to the compiled data and derived products with the community through a web page user interface.

Project Activities

The requested hardware shall be utilized to prepare a dataset that will be used to study and predict weather events that could potentially disrupt aviation and airport operations. Within the data science and machine learning communities, Physics Informed Machine Learning (PIML) is an emerging discipline with a rapidly developing area of research that seeks to leverage the wealth of scientific and domain-specific knowledge to improve the effectiveness of machine learning models. In our study, we are developing a PIML model to predict weather conditions leveraging the fundamental laws that govern the atmosphere. The underlying idea or vision of PIML is to advance our scientific understanding by producing generalizable ML models that are scientifically consistent and interpretable. Therefore, through this project, we wish to convince the community of the application of PIML methods to weather prediction and consequently, advance the use of PIML in the field of aviation meteorology. Additionally, we wish to contribute to the PIML community by presenting our work on weather prediction as a use-case.

As part of the previous efforts at ERAU, a design, and prototype of an aviation big data warehouse were presented [1]. A data pipeline was developed by the team using state-of-the-art technologies to extract flight routes from the TFMS flight data. In the presented work, Apache Nifi and Python were used for data ingestion and pre-processing. Apache Hadoop was used to facilitate both the storage and processing of data. Finally, Apache Hive was used to support SQL-like queries on the data. As a part of another effort to study weather-induced flight delays [2], it was identified that fusing the Global Forecast System (GFS) data with the flight delay data resulted in the faster convergence of the deep learning (DL) model. In the past, Apache NiFi was used to build a data pipeline to process the TFMS data only. In the current study, Apache NiFi is being additionally used to ingest and process big data (METAR and GFS) from disparate data sources. Big data tools such as Hadoop Distributed File System (HDFS) and Apache Hive are being examined for handling such data. Therefore, through this study, we intend to advance the community's knowledge of the application, considerations, advantages, and challenges of using these tools.

Resources Requested

If fully funded, we will purchase two (2) HPE ProLiant DL385 Gen 10 Plus v2 systems. Each will have a single 32 core AMD Epyc 7003 7513 Dotriaconta-core CPU and eight 8GB DIMMs for 64 GB of memory. These machines have a potential upgrade path for dual CPU and four times as much memory. If funded, the AAS Department will evaluate existing internal funds that might be available for these upgrades. The department is also anticipating the purchase of one or more GPU cards to enhance these systems.

These new systems will be clustered using additional resources from other HPE ProLiant systems that are already part of the Meteorology Program's computing infrastructure. While the initial use of this hardware will be for the processing of the previously described datasets, it is envisioned that the systems will also be available for other Machine Learning applications that colleagues are working on.

Information Technology Support Available

The Meteorology Program has a fairly robust computing environment that at one time supported Linux computing for all of our campus. Our IT infrastructure now supports home directories for both faculty and students. Many labs on campus have computers with dual-boot configurations, offering either RedHat Linux or MS Windows environments. Remote Linux desktops for all campus users are provided through the Meteorology Program infrastructure.

The Meteorology Program has six HPE ProLiant systems, and three older Dell systems, that provide data processing, filesharing, and hosting of various virtual machines (VM), including a cluster of VMs that are used by faculty and students for remote Linux computing via SSH or Remote Desktop protocols. The

resources also include over 200 TB of storage space, with plans to more than double this in the near future. Data are shared internally using the Unidata LDM, with two NOAAPort receiver systems to provide access to all the data needed to populate our EDEX server for the AWIPS CAVE application, which is used in multiple courses that we offer. We have regular data processing that generates images for our web server, using the Unidata provided GEMPAK and McIDAS software packages. We use the MetPy Python environment for our *WX 272 Meteorological Instruments and Data Analysis* course. We also participate in the Unidata IDD, receiving CONDUIT and GOES satellite data from the IDD and ADDE services. We have plans to participate in a distributed IDD effort with Unidata, taking advantage of the fact that we have different equinox satellite eclipse times from western locations.

Benefits for Education or Research

The requested systems will add significant computing resources to our infrastructure, as many of the existing systems are operating near their limits. The systems will be accessible to other faculty and students that have a need for these tools. As an example, a colleague is working on a Machine Learning project to help predict tornadic supercell events that also have flash flood warnings at the same time.

Potential Community Benefit

The dataset being prepared contains information about all flights that were impacted by one or more TMIs because of inclement weather at or around the airport, or at the destination airport. This includes GDP and GS at the airport due to bad weather conditions (including but not limited to thunderstorms, snow/ice, wind, low ceiling, etc.), flight rerouting, airspace flow program, and deicing operations at the airport. The weather conditions at the airport are captured from the airport configuration messages (APTC) of the TFMS data. This includes weather conditions like visual meteorological conditions (VMC), Instrument meteorological conditions (IMC), Low VMC (LVMC), and Low IMC (LIMC). Additionally, weather observations from the ASOS data have been merged with the TFMS dataset for detailed weather conditions at the airports. Further, the forecasts from the Global Forecast System (GFS) are to be added to this dataset. To the best of our knowledge, such a dataset is not available publicly.

This dataset shall be made available to the Unidata community through one or more Internet protocols. At a minimum, forecasts, and access to the database will be accessible through a website. If there is community interest, then the forecasts could be distributed through our outward facing LDM server.

Broadening the scope of the Unidata community

The dataset being prepared has detailed information about flights, including scheduled and actual times of arrivals and departures, taxi times, actual wheels off and on times, the weather conditions at the airports during their scheduled and actual departure/arrival times, and any implemented TMIs. It is anticipated that this dataset would be potentially useful to others in the community studying weather impacts on aviation. It is expected that the dataset would benefit both the aviation and meteorology communities, as the system will accumulate data over time and future data mining opportunities will only increase.

Contribution to education & research

From our literature review, it has been observed that using machine learning for predicting flight delays has been studied widely. However, the dataset most widely used in such studies is being obtained from the Bureau of Transportation Statistics (BTS). The BTS data provides a post-processed record for airlines participating in BTS data collection. The data are reported by the airlines, and the assignment of weather impacts is not always accurate. This dataset is not made available to the researchers in real-time, nor is it available in near real-time. For example, at the time of this proposal submission (March 2023) the latest data provided by the BTS is December 2022. On the other hand, TFMS data is made available to ERAU as it is recorded, and TFMS provides near real-time flight, flow information of the airspace, and weather conditions at the airports. The TFMS dataset is expected to be significantly more helpful to the researchers studying flight delays.

Additionally, to study weather impacts on aviation, a fusion of aviation and meteorological data is required. Literature shows that for every study, flight data collected from BTS must be pre-processed as the data are recorded in local time for every airport. TFMS data on the other hand is recorded using a UTC time reference. Therefore, using the proposed dataset would save the researchers effort and time. Further, the availability of such a dataset is expected to foster research in areas like predicting the nature and duration of TMIs, correlations in such decisions with observations and forecasts, and other yet to be conceived questions that could be answered.

Budget

The attached quote from the ERAU IT preferred vendor, Agilant Solutions Inc, is for an HPE ProLiant server that is consistent with much of our existing hardware. We are requesting two of these systems.

- HPE ProLiant DL385 Gen10 Plus v2 8SFF Configure-to-order Server $9,992.40
- HPE ProLiant DL385 Gen10 Plus v2 8SFF Configure-to-order Server $9,992.40

Total Requested Funds $19,984.80

The requested hardware has been identified as a suitable solution to improve our data processing. This configuration was confirmed through discussions with developers of the Apache NiFi data processing software that we are using.

Installation of the hardware will be conducted by the Applied Aviation Sciences Systems Administrator with assistance from our campus IT personnel if needed. We have ample space on our 10 Gb and 1 Gb networking switches, so integration should not be a problem. The installation of the operating system and data analysis tools will be done by the PIs, graduate students, and Sys. Admin. The timing of this initial effort should coincide with our summer semester when teaching conflicts will not be an issue for time. There Embry-Riddle Aeronautical Universities' federally negotiated on-campus rate excludes capital equipment from its facilities and administrative rate.

Project Milestones

Task	Time
Task (a): Procure equipment	4 weeks
Task (b): Hardware installation	2 weeks
Task (c): Software installation	1 week
Task (d): Data preparation using new computing resources	4 weeks
Task (e): Demonstration of AI/ML model training and testing	20 weeks
Task (f): Dissemination of results via website	4 weeks
Task (g): Product technical report	2 weeks

Selected task details

Task (c): Software Installation: After the required hardware has been procured and installed, setting up the software environment shall begin immediately. Based on our experience of configuring a similar environment on the current resources, this task shall be straightforward.

Task (d): Data Preparation using new computing resources: As aforementioned, in the current setup it takes approximately 4 days to process a month's worth of TFMS data. It takes additional days when GFS data is processed along with TFMS. However, with the requested hardware it is expected to make processing this data faster. We intend to start with 2022 TFMS data. However, to include seasonality we will include additional years of TFMS, METAR, and forecast data.

Task (e): Demonstration of AI/ML model training and testing: As mentioned above, the dataset prepared using the requested hardware shall be used as input for the PIML model to predict weather conditions that would interrupt the airport and aviation operations. Depending on the size of the data, it would take hours to train and test the model every time, with additional time for model tuning. This shall also include time for performing literature reviews from time to time.

Task (g): Product technical report: Once we have the model results, we intend to share detailed information about the computing resources utilized, dataset prepared, PIML model, results, and lessons learned with the community in a detailed report.

References

[1] Thota, Ravi & Bawa, Gurvir & Stansbury, Richard. (2020). Design and Prototyping of an Aviation Big Data Repository. 10.2514/6.2020-0319.

[2] Gurvir K. Bawa, Richard S. Stansbury and Christopher G. Herbster. "A Study of Weather-Induced Flight Delay using Deep Learning Techniques," AIAA 2022-4080. AIAA AVIATION 2022 Forum. June 2022.

Agilant Solutions Inc
Miami
Steve Perea
6303 Blue Lagoon DriveSuite 400
Miami, FL 33126

Phone: 305-666-6804

QUOTATION

#040475 v1
Quote Date: 3.13.2023
Expires: 4.10.2023

BILL TO	SHIP TO
Embry-Riddle Aeronautical University 1 Aerospace Blvd Daytona, AZ 86301 United States	**Embry-Riddle Aeronautical University** 1 Aerospace Blvd Daytona, AZ 86301 United States

CUSTOMER NUMBER	PROJECT NAME	TERMS
131	AI Server 32/Core	Net 30
CUSTOMER NAME	**CUSTOMER PHONE**	**ACCOUNT EXECUTIVE**
Daniel Stubbs	13862267172	Steve Perea

AI 32 Core/128GB

Part Number	Description	Tax	Qty	Unit Price	Extended Price
P348411-B21	**HPE ProLiant DL385 Gen10 Plus v2 8SFF Configure-to-order Server**	N	1	$1,355.69	$1,355.69
P38684-B21	**HPE AMD EPYC 7003 7513 Dotriaconta-core (32 Core) 2.60 GHz Processor Upgrade - 128 MB L3 Cache - 3.65 GHz Overclocking Speed - Socket SP3 - 200 W - 64 Threads**	N	1	$2,711.00	$2,711.00
P07638-B21	**HPE SmartMemory 8GB DDR4 SDRAM Memory Module - For Server - 8 GB (1 x 8GB) - DDR4-3200/PC4-25600 DDR4 SDRAM - 3200 MHz - CL22 - 1.20 V - Registered - 240-pin - DIMM**	N	8	$129.62	$1,036.96
P28586-B21	**HPE 1.20 TB Hard Drive - 2.5" Internal - SAS (12Gb/s SAS) - Server Device Supported - 10000rpm**	N	2	$234.91	$469.82
P26259-B21	**HPE Broadcom BCM57412 Ethernet 10Gb 2-port SFP+ Adapter for HPE - PCI Express 3.0 x8 - 1.25 GB/s Data Transfer Rate - 2 Port(s) - Optical Fiber - 10GBase-X - SFP+ - Standup**	N	1	$459.56	$459.56
P01366-B21	**HPE Battery - For RAID Controller - Battery Rechargeable**	N	1	$65.09	$65.09
P26279-B21	**HPE Broadcom MegaRAID MR416i-a SAS Controller - 12Gb/s SAS - PCI Express 4.0 x16 - Plug-in Module - RAID Supported - 0, 1, 5, 6, 10, 50, 60 RAID Level - 16 Total SAS Port(s) - 16 SAS Port(s) Internal - PC - 4 GB**	N	1	$740.79	$740.79
P08449-B21	**HPE Ethernet 1Gb 4-port BASE-T I350-T4 OCP3 Adapter - PCI Express 2.0 x4 - 4 Port(s) - 4 - Twisted Pair - 1000Base-T - Plug-in Card**	N	1	$468.36	$468.36
455883-B21	**HP BladeSystem c-Class 10Gb Short Range Small Form-Factor Pluggable Option**	N	2	$327.44	$654.88
P14608-B21	**HPE DL38X Gen10 Plus Maximum Performance Fan Kit**	N	1	$120.53	$120.53
P38995-B21	**HPE 800W Flex Slot Platinum Hot Plug Low Halogen Power Supply Kit - Hot-pluggable - 96% Efficiency**	N	2	$244.53	$489.06
BD505A	**HPE Integrated Lights-Out Advanced - Subscription License - 1 Server License - 3 Year - Standard**	N	1	$192.81	$192.81

Agilant Solutions Inc
Miami
Steve Perea
6303 Blue Lagoon DriveSuite 400
Miami, FL 33126

Phone: 305-666-6804

QUOTATION

#040475 v1
Quote Date: 3.13.2023
Expires: 4.10.2023

AI 32 Core/128GB

Part Number	Description	Tax	Qty	Unit Price	Extended Price
P55467-B21	**HPE ProLiant DL38x 8SFF SAS/SATA Tri-Mode Cable Kit**	N	1	$31.03	$31.03
P13771-B21	**HPE Trusted Platform Module 2.0 Gen10 Plus Black Rivets Kit - Black**	N	1	$40.53	$40.53
P14610-B21	**HPE DL38X Gen10 Plus High Performance Heat Sink Kit - Riser Card**	N	1	$110.93	$110.93
P22018-B21	**HPE DL38X Gen10 Plus 2U SFF Easy Install Rail Kit - For Server - 2U Rack Height - Rack-mountable**	N	1	$51.20	$51.20
HU4B2A3	**HPE 3Y Tech Care Basic SVC**	N	1	$0.00	$0.00
HU4B2A3#ZSF	**HPE Proliant DL385 Gen10 Plus V2 Support**	N	1	$994.16	$994.16

QUESTIONS? PLEASE EMAIL YOUR ACCOUNT EXECUTIVE. AGILANT SOLUTIONS, INC., A NY CORPORATION, FEIN 20-3074058

Subtotal	**$9,992.40**
Sales Tax	**$0.00**
Total	**$9,992.40**

Thank you! I appreciate your trust and interest in allowing Agilant Solutions, Inc. this opportunity to assist with you IT projects, plans and initiatives.

Note:
Certain manufacturers have implemented restrictive return policies that prevent Agilant from being able to accept returns or offer exchanges, replacements or credits on their products – for any reason. For these products, all support as well as exchanges or replacements for defective merchandise (DOA), including within the first 15 days from date of shipment, must be handled directly with the product manufacturer. If you are unsure if your quote/order contains items that are subject to restricted returns, please consult with your Account Executive.

All price quotations are subject to Agilant Terms & Conditions of sale. Individual requirements, options and custom requests will impact price and terms.

All information contained within this quotation is confidential to Agilant Solutions, Inc. and its intended parties.

NASPO

Description	Qty
NASPO ValuePoint (43211500-WSCA-15-ACS)	

Quote Summary	Amount
AI 32 Core/128GB	$9,992.40
Total:	**$9,992.40**

Agilant Solutions Inc
Miami
Steve Perea
6303 Blue Lagoon DriveSuite 400
Miami, FL 33126

Phone: 305-666-6804

QUOTATION

#040475 v1
Quote Date: 3.13.2023
Expires: 4.10.2023

Acceptance

Miami

Steve Perea
Signature / Name

03/13/2023
Date

Embry-Riddle Aeronautical University

Daniel Stubbs
Signature / Name Initials

Date