Cybersecurity & Information Systems Information Analysis Center

# CSIAC TECHNICAL INQUIRY (TI) RESPONSE REPORT

## Artificial Intelligence/Machine Learning (AI/ML) Tools for the Repository for Analytics in a Virtualized Environment (RAVEn) Platform

**Report Number:**

CSIAC-BCO-2025-672

**Completed April 2025**


**CSIAC** is a U.S. Department of Defense Information Analysis Center

**MAIN OFFICE**
4695 Millennium Drive
Belcamp, MD 21017-1505
Office: 443-360-4600

**REPORT PREPARED BY:**
Joseph M. Friar
Office: CSIAC





CSIAC is sponsored by the Defense Technical Information Center, with policy oversight provided by the Office of the Under Secretary of Defense for Research and Engineering. CSIAC is operated by the SURVICE Engineering Company.

# REPORT DOCUMENTATION PAGE

*Form Approved*
*OMB No. 0704-0188*



| **1. REPORT DATE** *(DD-MM-YYYY)* | **2. REPORT TYPE** | **3. DATES COVERED** *(From – To)* |
|---|---|---|
| April 2025 | Technical Research Report | |

**4. TITLE AND SUBTITLE**

Artificial Intelligence/Machine Learning (AI/ML) Tools for the Repository for Analytics in a Virtualized Environment (RAVEn) Platform

**5a. CONTRACT NUMBER**
FA8075-21-D-0001

**5b. GRANT NUMBER**

**5c. PROGRAM ELEMENT NUMBER**

**6. AUTHOR(S)**

Joseph M. Friar

**5d. PROJECT NUMBER**

**5e. TASK NUMBER**

**5f. WORK UNIT NUMBER**

**7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)**

Cybersecurity & Information Systems Information Analysis Center (CSIAC)
SURVICE Engineering Company
4695 Millennium Drive
Belcamp, MD 21017-1505

**8. PERFORMING ORGANIZATION REPORT NUMBER**

CSIAC-BCO-2025-672

**9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)**

Defense Technical Information Center (DTIC)
8725 John J. Kingman Road
Fort Belvoir, VA 22060-6218

**10. SPONSOR/MONITOR'S ACRONYM(S)**

**11. SPONSOR/MONITOR'S REPORT NUMBER(S)**

**12. DISTRIBUTION/AVAILABILITY STATEMENT**

Distribution Statement A. Approved for public release: distribution is unlimited.

**13. SUPPLEMENTARY NOTES**

**14. ABSTRACT**

Artificial intelligence (AI) and machine learning (ML) are transforming the ways organizations interact with data and generate insights from it. This report introduces key concepts present using AI for data analytics and highlights the most recent tools and platforms in the field. Several commercial, government, or open-source software tools were identified during research. These tools could be used to leverage AI/ML to support the inquirer in data-analytic efforts relating to the Repository for Analytics in a Virtualized Environment (known as RAVEn) platform.

**15. SUBJECT TERMS**
artificial intelligence, machine learning, data analytics

| **16. SECURITY CLASSIFICATION OF:** U | | | **17. LIMITATION OF ABSTRACT** | **18. NUMBER OF PAGES** | **19a. NAME OF RESPONSIBLE PERSON** |
|---|---|---|---|---|---|
| **a. REPORT** | **b. ABSTRACT** | **c. THIS PAGE** | UU | 17 | Ted Welsh, CSIAC Director |
| U | U | U | | | **19b. TELEPHONE NUMBER** *(include area code)* 443-360-4600 |

**Standard Form 298 (Rev. 8-98)**
**Prescribed by ANSI Std. Z39.18**

# About

## DTIC and CSIAC

The Defense Technical Information Center (DTIC) preserves, curates, and shares knowledge from the U.S. Department of Defense's (DoD's) annual multibillion-dollar investment in science and technology, multiplying the value and accelerating capability to the Warfighter. DTIC amplifies this investment by collecting information and enhancing the digital search, analysis, and collaboration tools that make information widely available to decision-makers, researchers, engineers, and scientists across the Department.

DTIC sponsors the DoD Information Analysis Centers (DoDIAC), which provide critical, flexible, and cutting-edge research and analysis to produce relevant and reusable scientific and technical information for acquisition program managers, DoD laboratories, Program Executive Offices, and Combatant Commands. The IACs are staffed by, or have access to, hundreds of scientists, engineers, and information specialists who provide research and analysis to customers with diverse, complex, and challenging requirements.

The Cybersecurity & Information Systems Information Analysis Center (CSIAC) is a DoDIAC sponsored by DTIC to provide expertise in four technical focus areas: cybersecurity, knowledge management & information sharing, modeling & simulation, and software data & analysis. CSIAC is operated by SURVICE Engineering Company under contract FA8075-21-D-0001.

## TI Research

A chief service of the DoDIAC is free technical inquiry (TI) research limited to four research hours per inquiry. This TI response report summarizes the research findings of one such inquiry. Given the limited duration of the research effort, this report is not intended to be a deep, comprehensive analysis but rather a curated compilation of relevant information to give the reader/inquirer a "head start" or direction for continued research.

# Abstract

Artificial intelligence (AI) and machine learning (ML) are transforming the ways organizations interact with data and generate insights from it. This report introduces key concepts present using AI for data analytics and highlights the most recent tools and platforms in the field. Several commercial, government, or open-source software tools were identified during research. These tools could be used to leverage AI/ML to support the inquirer in data-analytic efforts relating to the Repository for Analytics in a Virtualized Environment (known as RAVEn) platform.

# Contents



# List of Tables

# 1.0 TI Request

## 1.1 Inquiry

Can artificial intelligence (AI)/machine learning (ML) tools that can perform analysis on a wide range of data that is centralized in the Repository for Analytics in a Virtualized Environment (RAVEn) platform be identified?

## 1.2 Description

The Cybersecurity and Information Systems Information Analysis Center (CSIAC) was asked to provide information on AI/ML tools that can be used to aid in data-analytic activities. Tools that could be implemented into the RAVEn platform to enhance its capabilities were of particular interest, but some related software that could prove helpful in the inquirer's mission has been included as well.

# 2.0 TI Response

RAVEn is a data repository that supports the U.S. Department of Homeland Security (DHS) mission to enforce and investigate violations of U.S. criminal, civil, and administrative laws. DHS reached out to CSIAC for a list of AI/ML tools that can analyze data in the repository, enhance entity resolution, and better streamline data insights on the department level. According to the DHS [1]:

> The United States Immigration and Customs Enforcement (ICE) Homeland Security Investigations (HSI) Innovation Lab is developing an analytical platform called the Repository for Analytics in a Virtualized Environment (RAVEn). RAVEn will facilitate large, complex analytical projects to support ICE's mission to enforce and investigate violations of U.S. criminal, civil, and administrative laws. RAVEn also enables users to develop new tools to analyze trends and isolate criminal patterns as HSI mission needs arise.

CSIAC performed an open-source search to find existing software tools that can leverage AI/ML to help process and fuse data that will be utilized by the RAVEn platform. While the RAVEn platform deals with a specific source of data and types of information, these tools should be able to assist with analytical work at HSI.

## 2.1 AI in Data Analytics

From this research, several commercial, government, or open-source software tools were identified. These tools could be used to leverage AI/ML to enhance and support data-analytic efforts relating to the RAVEn platform. AI-based analytics platforms automate data classification and pattern detection to facilitate the transformation of raw data into valuable intelligence [2].

### 2.1.1 Natural Language Processing (NLP) for Unstructured Data

NLP is being used to understand different types of unstructured data such as emails, support logs, and socialmedia messages. A notable feature of these technologies can be found in the ability to "understand the language of your business" [3] or system and tailor its functions to be more specific to the project at hand. International Business Machines (IBM) Watson Natural Language Understanding and Amazon Comprehend are technologies that use AI to process and interpret text data, which allow users to interact with many text-based data sources and interpret them [3, 4].

### 2.1.2 Automated Machine Learning (AutoML)

Platforms such as DataRobot, H2O.ai, and Google Cloud AutoML are AutoML applications that improve the model development process by automating data preprocessing, model selection, and parameter tuning [5]. The tools that are a part of the package enable business users and analysts who do not have deep programming or ML skills to conduct advanced analytics and make predictions with their data. The tools also provide users with model explainability, which ensures transparency and supports the model governance frameworks.

### 2.1.3 AI-Driven Data Preparation and Integration

Before being used, data often need to be "prepared" by cleaning, organizing, and getting rid of duplicate entries, among many other small details. Talend and Trifacta are tools that use ML to identify these anomalies and help prepare the dataset for the next steps [6]. This way, the manual effort needed for these tasks is reduced, accelerating the data preparation process.

### 2.1.4 Real-Time Analytics and Streaming Data

Apache Kafka and Google Cloud Dataflow are real-time analytics platforms that are integrating AI models to carry out the processing of streaming data in applications by using fraud detection, predictive maintenance, and real-time customer interaction. This set of technologies allows for quick reactions in response to the analytics being displayed, improving operational agility [7].

### 2.1.5 Model Monitoring and Governance

The usage of AI in analytics depends on the accuracy of the models, which can be improved with the proper monitoring and governance systems in place. Tools such as MLflow, Fiddler.AI, and Evidently AI are enablers that help by providing automated metric tracking and logging of model behavior throughout the life cycle, ensuring the compliance with the future regulatory needs on transparency and accountability in AI systems [8].

### 2.1.6 Predictive and Prescriptive Analytics

ML and AI-powered statistical models for predictive analytics are used to collectively analyze past performance with the aim of predicting future outcomes, helping in tasks such as data planning and operational risk measurement. Prescriptive analytics expand these capabilities via the simulation and optimization of the model by suggesting actions. For example, reinforcement learning has become more common in inventory management and logistics for making better dynamic decision-making possible [9].

### 2.1.7 AI-Augmented Business Intelligence (BI) Tools

BI platforms, such as Microsoft Power BI, Tableau, and Qlik, use AI, facilitating everyday business-related functions. These include features such as natural language query, intelligent data discovery, and auto visualizations. In this way, these platforms facilitate a more user-friendly interaction with data for the nontechnical crowd, providing freedom from a dependency on data science teams. The ability of a question-and-answer (better known as Q&A) feature in Power BI, for example, allows users to ask questions in a natural language format [10]. Explain Data from Tableau is a feature that explains and describes the statistical model trends and outliers by statistical methods [11].

## 2.2 Large Language Models (LLMs) in Data Analytics

An Information Analysis Center team of in-house data scientists reviewed this technical inquiry and suggested three options for an LLM:

1. Cloud provider to run a controlled LLM
2. LLM application programming interface (API), which is easier to set up but offers less flexibility
3. Local, controlled LLM, which can be more cost effective

The choice depends on scale and budget, and different models can be compared in the Hugging Face Model Leaderboard [12]. For fine-tuning, CSIAC data scientists suggested

something in the Llama family, as it typically performs well as an “open-source” model (there is a license that allows for up to a million users). If databases are being used, the user can utilize a retrieval-augmented generation (RAG) approach. However, if data includes numbers, a custom RAG system with several intermediate steps may be needed. Models generally struggle with numerical responses, so one might have to search the database for numbers, convert them into text fields, and then combine that text with other inputs to create a cohesive response.

Some general key words that may help users find an appropriate LLM to fit their needs are:

- Retrieval-Augmented Generation, or RAG
- Langchain
- Hugging Face

Both open-source and custom models allow developers to “plug and play” the data with an ML construction and deployment platform. Something like Azure Machine Learning, Google Cloud AI Platform, PyTorch Ecosystem tools, or BigML may be helpful before deploying and monitoring the model with that provider’s infrastructure. If an individual is looking to build a custom ML system with multiple steps or wanting to have more control than what proprietary options allow, selecting an LLM to build into a process like RAG might be worth looking into.

## 2.3 Leading Tools and Platforms

Table 1 details various tools and platforms for AI-powered data analytics, along with a description of each. Each description has a reference citation through which additional information can be found.

## Table 1. Leading Tools and Platforms for AI-Powered Data Analytics

| Name | Category | Source Description |
|---|---|---|
| Amazon SageMaker | AutoML and Machine Learning Operations (MLOps) Platforms | "Bringing together widely adopted [Amazon Web Services] AWS ML and analytics capabilities, Amazon SageMaker delivers an integrated experience for analytics and AI with unified access to all your data" [13]. |
| DataRobot | AutoML and MLOps Platforms | "DataRobot delivers the industry-leading agentic AI applications and platform that maximize impact and minimize risk for your business" [14]. |
| Google Cloud: Vertex AI Studio | AutoML and MLOps Platforms | "Streamline your foundation model workflows with Vertex AI Studio. Rapidly prototype, refine, and seamlessly deploy models to your applications" [15]. |
| Google Cloud: AutoML | AutoML and MLOps Platforms | "Train high-quality custom ML models with minimal effort and ML expertise" [16]. |
| H2O.ai | AutoML and MLOps Platforms | "H2O.ai brings model risk management (MRM) to GenAI, combining enterprise-grade model evaluation with h2oGPTe and H2O Model Validation to make AI measurable, explainable, and compliant for banks and financial services" [17]. |
| Microsoft Azure Machine Learning | AutoML and MLOps Platforms | "Streamline prompt engineering and ML model workflows. Accelerate model development with powerful AI infrastructure" [18]. |
| Microsoft Power BI | BI and Visualization Tools | "Turn your data into visuals with advanced data-analysis tools, AI capabilities, and a user-friendly report-creation tool" [19]. |
| Tableau | BI and Visualization Tools | "Unlock your data's full potential with agentic analytics. Deliver trusted insights for everyone and enrich agents with Tableau Next, built on Salesforce and integrated with Agentforce" [20]. |
| Databricks | Data Warehousing and Big Data Platforms | "The Databricks Data Intelligence Platform allows your entire organization to use data and AI. It's built on a 'lakehouse' [21] to provide an open, unified foundation for all data and governance and is powered by a data intelligence engine that understands the uniqueness of your data" [22]. |
| Google BigQuery | Data Warehousing and Big Data Platforms | "BigQuery is the autonomous data to AI platform, automating the entire data life cycle, from ingestion to AI-driven insights, so you can go from data to AI to action faster" [23]. |
| Snowflake | Data Warehousing and Big Data Platforms | "Snowflake brings AI to the governed data, enabling teams to run analytical workflows on unstructured data, develop agentic apps, and train models using both structured and unstructured data—all with minimal operational overhead and end-to-end governance" [24]. |
| IBM Watson Studio | Enterprise Analytics Tools | "IBM Watson Studio empowers data scientists, developers, and analysts to build, run, and manage AI models and optimize decisions anywhere on IBM Cloud Pak for Data [25]. Unite teams, automate AI lifecycles, and speed time to value on an open multicloud architecture" [26]. |
| SAS | Enterprise Analytics Tools | "With our data and AI platform, you can understand what's happening with your data now, predict how to pivot seamlessly, and get results faster" [27]. |

| Name | Category | Source Description |
|---|---|---|
| Ask Sage | Generative AI and LLM Platforms | "Thanks to Ask Sage connectors to data lakes and APIs, Ask Sage enables seamless access to real-time data and brings data-analysis capability to the next level" [28]. |
| ChatGPT | Generative AI and LLM Platforms | "We've trained a model called ChatGPT that interacts in a conversational way. The dialogue format makes it possible for ChatGPT to answer follow-up questions, admit its mistakes, challenge incorrect premises, and reject inappropriate requests" [29]. |
| Meta Llama | Generative AI and LLM Platforms | "The most accessible and scalable generation of Llama is here. Native multimodality, mixture-of-experts models, superlong context windows, step changes in performance, and unparalleled efficiency. All in easy-to-deploy sizes custom fit for how you want to use it" [30]. |
| BigML | Low-Code/No-Code AI and Analytics Tools | "BigML provides a selection of robustly engineered ML algorithms proven to solve real-world problems by applying a single, standardized framework across your company" [31]. |
| RapidMiner | Low-Code/No-Code AI and Analytics Tools | "Altair RapidMiner is a powerful data-analytics and AI platform that connects siloed data, unlocks hidden insights, and accelerates innovation with advanced analytics and AI-driven automation" [32]. |
| Keras 3 | ML and Deep Learning Frameworks | "As a multiframework API, Keras can be used to develop modular components that are compatible with any framework—JAX, TensorFlow, or PyTorch" [33]. |
| PyTorch | ML and Deep Learning Frameworks | "Scalable distributed training and performance optimization in research and production is enabled by the torch-distributed backend" [34]. |
| TensorFlow | ML and Deep Learning Frameworks | "TensorFlow makes it easy to create ML models that can run in any environment. Learn how to use the intuitive APIs through interactive code samples" [35]. |
| JAX | ML and Deep Learning Frameworks | "JAX is a Python library for accelerator-oriented array computation and program transformation, designed for high-performance numerical computing and large-scale ML" [36]. |
| Python | Programming Languages | "Python is developed under an [Open-Source Initiative] OSI-approved open-source license, making it freely usable and distributable, even for commercial use. Python's license is administered by the Python Software Foundation [37]" [38]. |
| R | Programming Languages | "R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of [Uniplexed Information Computing System] UNIX platforms, Windows and MacOS" [39]. |
| SQL | Programming Languages | "Structured query language (SQL) is a programming language for storing and processing information in a relational database. A relational database stores information in tabular form, with rows and columns representing different data attributes and the various relationships between the data values" [40]. |

| Name | Category | Source Description |
|---|---|---|
| Apache Kafka | Real-Time Analytics and Streaming Tools | “Apache Kafka is an open-source distributed-event streaming platform used by thousands of companies for high-performance data pipelines, streaming analytics, data integration, and mission-critical applications” [41]. |

# References

1. U.S. Department of Homeland Security. "DHS/ICE/PIA-055 Repository Analytics and Virtualized Environment (RAVEn)." Homeland Security, https://www.dhs.gov/publication/dhsicepia-055-repository-analytics-virtualized-environment-raven, accessed on 29 April 2025.

2. Davenport, T. H., and R. Ronanki. "Artificial Intelligence for the Real World." Harvard Business Review, https://hbr.org/2018/01/artificial-intelligence-for-the-real-world, accessed on 22 April 2025.

3. International Business Machines Corporation. "Watson Natural Language Understanding." IBM, https://www.ibm.com/cloud/watson-natural-language-understanding, accessed on 22 April 2025.

4. Amazon Web Services, Inc. "Amazon Comprehend." AWS, https://aws.amazon.com/comprehend/, accessed on 22 April 2025.

5. Zöller, M.-A., and M. F. Huber. "Benchmark and Survey of Automated Machine Learning Frameworks." *Journal of Artificial Intelligence Research*, vol. 70, pp. 409–472, 26 January 2021.

6. Trifacta. "Data Wrangling With AI and Machine Learning." https://www.trifacta.com/resources/whitepapers, accessed on 22 April 2025.

7. Google LLC. "Dataflow: Real-Time and Batch Data Processing." Google Cloud, https://cloud.google.com/dataflow, accessed on 22 April 2025.

8. Sculley, D., G. Holt, D. Golovin, E. Davydov, T. Phillips, D. Ebner, V. Chaudhary, M. Young, J. F. Crespo, and D. Dennison. "Hidden Technical Debt in Machine Learning Systems." NeurIPS Proceedings, https://proceedings.neurips.cc/paper_files/paper/2015/file/86df7dcfd896fcaf2674f757a2463eba-Paper.pdf, 2015.

9. Bertsimas, D., and N. Kallus. "From Predictive to Prescriptive Analytics." *Management Science*, vol. 66, no. 3, pp. 1025–1044, March 2020.

10. Microsoft Corporation. "Power BI AI Capabilities." Learn, https://learn.microsoft.com/en-us/power-bi/ai-overview, accessed on 22 April 2025.

11. Salesforce, Inc. "Augmented Analytics: Explain Data." Tableau, https://www.tableau.com/products/new-features/explain-data, accessed on 22 April 2025.

12. Hugging Face. "Open LLM Leaderboard (Archived)." Spaces, https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard#/, accessed April 2025.

13. Amazon Web Services, Inc. "Amazon SageMaker." AWS, https://aws.amazon.com/sagemaker/, accessed April 2025.

14. DataRobot, Inc. "Why DataRobot?" DataRobot, https://www.datarobot.com/, accessed April 2025.

15. Google LLC. "Vertex AI Studio." Google Cloud, https://cloud.google.com/generative-ai-studio, accessed April 2025.

16. Google LLC. "AutoML." Google Cloud, https://cloud.google.com/automl, accessed April 2025.

17. H2O.ai. "The World's Best Deep Research." https://h2o.ai/, accessed April 2025.

18. Microsoft Corporation. "Azure Machine Learning." Microsoft: Azure, https://azure.microsoft.com/en-us/products/machine-learning, accessed April 2025.

19. Microsoft Corporation. "Power BI." Microsoft: Power Platform, https://www.microsoft.com/en-us/power-platform/products/power-bi/, accessed April 2025.

20. Salesforce, Inc. "Tableau Conference." Tableau From Salesforce, https://www.tableau.com/, accessed April 2025.

21. Databricks, Inc. "Data Lakehouse Architecture." Databricks, https://www.databricks.com/product/data-lakehouse, accessed April 2025.

22. Databricks, Inc. "Your Data. Your AI. Your Future." Databricks, https://www.databricks.com/, accessed April 2025.

23. Google, LLC. "BigQuery." Google Cloud, https://cloud.google.com/bigquery, accessed April 2025.

24. Snowflake, Inc. "Build the Future of AI and Apps." Snowflake, https://www.snowflake.com/en/, accessed April 2025.

25. International Business Machines Corporation. "IBM Cloud Pak for Data." IBM, https://www.ibm.com/products/cloud-pak-for-data, accessed April 2025.

26. International Business Machines Corporation. "IBM Watson Studio." IBM, https://www.ibm.com/products/watson-studio, accessed April 2025.

27. SAS Institute Inc. "Data and AI Solutions." SAS, https://www.sas.com/en_us/home.html, accessed April 2025.

28. Ask Sage, Inc. "Unleash the Power of Generative AI With Ask Sage." Ask Sage, https://www.asksage.ai/, accessed April 2025.

29. OpenAI, Inc. "Introducing ChatGPT." OpenAI, https://openai.com/index/chatgpt/, 30 November 2022.

30. Meta. "Llama 4: Leading Intelligence. Unrivaled Speed and Efficiency." https://www.llama.com/, accessed April 2025.

31. BigML, Inc. "Machine Learning Made Beautifully Simple for Everyone." BigML, https://bigml.com/, accessed April 2025.

32. Altair Engineering Inc. "Altair RapidMiner: Transform Your Data Strategy." Altair, https://altair.com/altair-rapidminer, accessed April 2025.

33. Keras. "About Keras 3." https://keras.io/getting_started/about/, accessed April 2025.

34. PyTorch. "Get Started." https://pytorch.org/, accessed April 2025.

35. TensorFlow. "An End-to-End- Platform for Machine Learning." https://www.tensorflow.org/, accessed April 2025.

36. The JAX Authors. "JAX: High-Performance Array Computing." JAX, https://docs.jax.dev/en/latest/, accessed April 2025.

37. Python Software Foundation. "The Python Software Foundation Is an Organization Devoted to Advancing Open-Source Technology Related to the Python Programming Language." PSF, https://www.python.org/psf-landing/, accessed April 2025.

38. Python Software Foundation. “Python Is a Programming Language that Lets You Work Quickly and Integrate Systems More Effectively.” Python, https://www.python.org/, accessed April 2025.

39. The R Foundation. “The R Project for Statistical Computing.” R, https://www.r-project.org/, accessed April 2025.

40. Amazon Web Services, Inc. “What Is SQL (Structured Query Language?” AWS, https://aws.amazon.com/what-is/sql/, accessed April 2025.

41. Apache Software Foundation. “Apache Kafka.” https://kafka.apache.org/, Kafka, accessed April 2025.

# Biography

**Mr. Friar** works with the Cybersecurity & Information Systems Information Analysis Center team as a research inquiry analyst. He graduated from Stevenson University, acquiring a B.S. in computer information systems. He is currently working toward an M.S. in information technology, with a concentration in systems engineering. At the SURVICE Engineering Company, he performs in-depth research relating to technology fields such as cybersecurity and information systems. Mr. Friar also works with government clients to provide them with information-oriented solutions and answer their technical inquiries.