Developing a Roadmap towards the Next Generation of Smart Cyberinfrastructure
February 25-27, 2020, Hyatt Regency Crystal City, VA,

Intelligent Data Analytics Environment (IDAE) for High Performance Cyberinfrastructure

Salim Hariri
NSF I/UCRC Center for Cloud and Autonomic Computing
University of Arizona
Nsfcac.arizona.edu
hariri@email.arizona.edu

The current machine learning algorithms are mainly developed to run on sequential platforms. With the current exponential growth in data, and the large scale of ML applications, it is becoming critically important to reduce the execution time and improve the scalability of the ML algorithms. The main goal of this whitepaper is to highlight the main research challenges that must be addressed to allow high performance cyberinfrastructure to exploit the emerging big data programming paradigms (MapReduce), artificial intelligence, machine learning algorithms, and high performance platforms (parallel, distributed and clusters of GPUs).

Research Challenges

- ***How to handle large-scale dynamic and heterogeneous data streams?***
  - ***The main research problem is here is how to detect accurately the changes in data streams***
- ***How to develop an intelligent recommender system that tells the user the best Ml algorithm to use and how it should be configured?***
- ***How to apply parallel/distributed algorithms and big data programming tools to speedup ML computations?***
- ***How do you validate and benchmark the proposed approach on a wide range of scientific and engineering applications.***

In this white paper, we will discuss ongoing research activities at the University of Arizona to address these challenges. Figure 1 shows an environment to develop an Intelligent Data Analytics Environment (IDEA) and Figure 2 shows how to detect changes in the data streams being analyzed. Figure 3 shows an approach to adopt the Ml algorithm so it can model the recent changes in data streams. The first three figures can potentially address the first two research challenges.
Figure 4 shows a High Performance Machine Learning Framework (HPMLF) that can overcome the last two research challenges. The HPMLF will leverage Big Data analytics tools, MapRedue, parallel/distributed algorithms and high performance platforms (Parallel/distributed systems and GPU cluster).

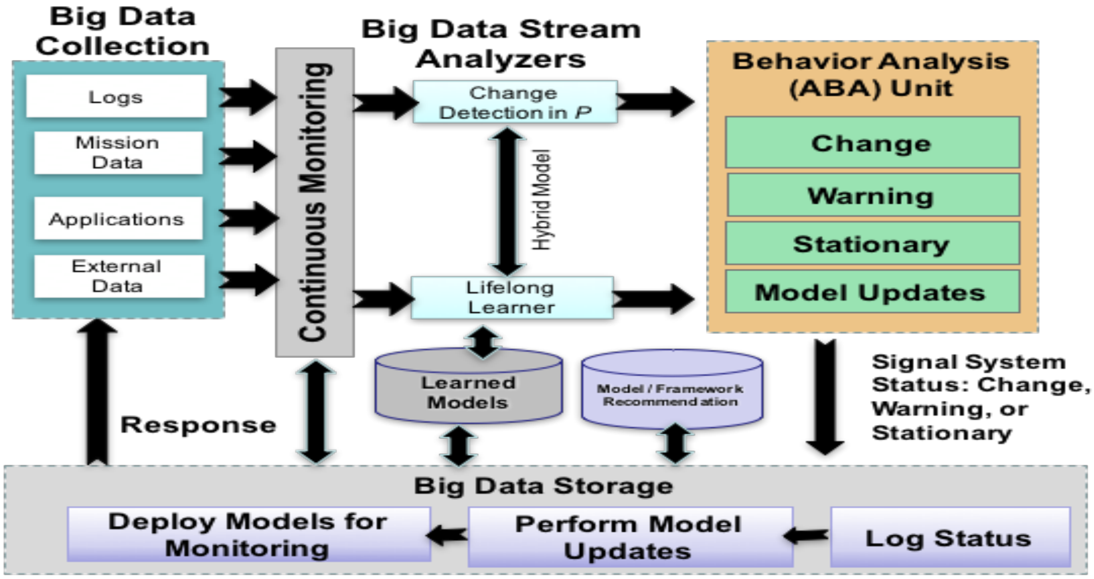
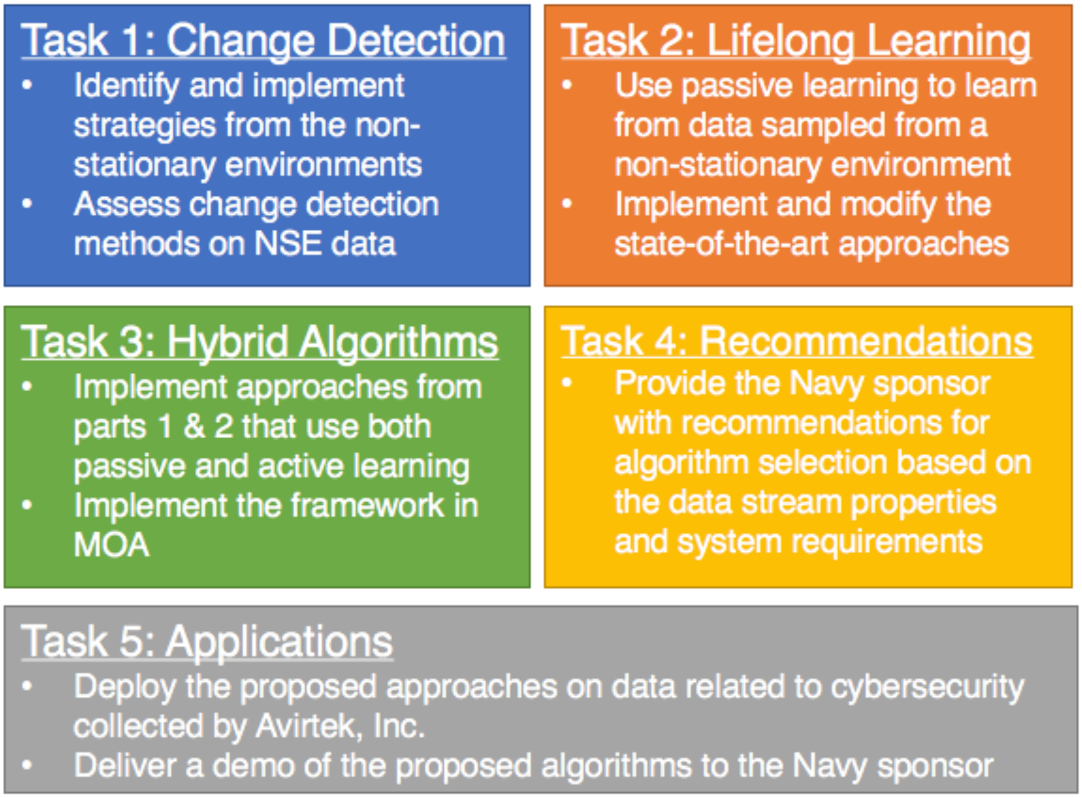


Figure 1. Intelligent data analytics environment.



Task 1: Change Detection

- Identify and implement strategies from the non-stationary environments
- Assess change detection methods on NSE data

Task 2: Lifelong Learning

- Use passive learning to learn from data sampled from a non-stationary environment
- Implement and modify the state-of-the-art approaches

Task 3: Hybrid Algorithms

- Implement approaches from parts 1 & 2 that use both passive and active learning
- Implement the framework in MOA

Task 4: Recommendations

- Provide the Navy sponsor with recommendations for algorithm selection based on the data stream properties and system requirements

Task 5: Applications

- Deploy the proposed approaches on data related to cybersecurity collected by Avirtek, Inc.
- Deliver a demo of the proposed algorithms to the Navy sponsor

Figure 2. Change detection tasks and applications.

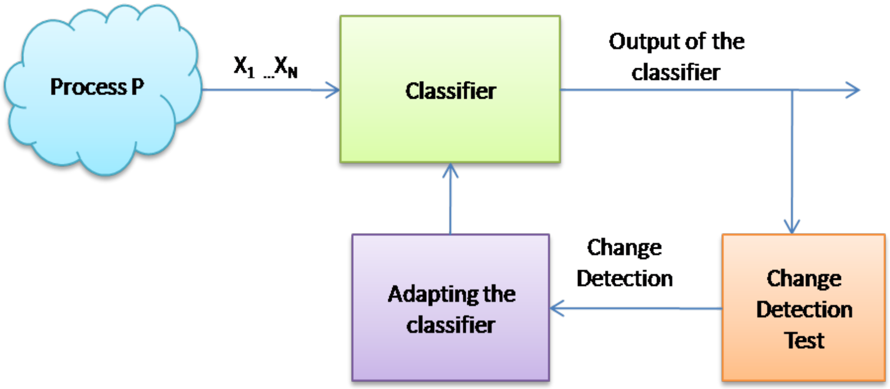


Figure 3. Real-time adaptive machine learning algorithm.

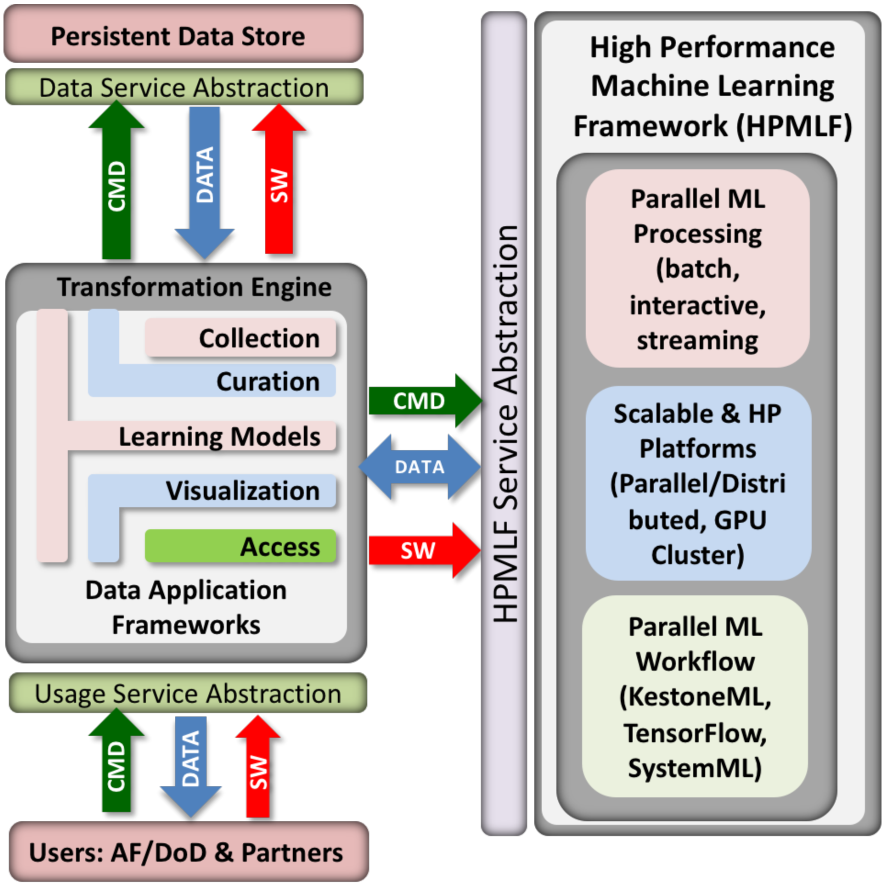

Persistent Data Store
Data Service Abstraction
CMD
DATA
SW
Transformation Engine
Collection
Curation
Learning Models
Visualization
Access
Data Application Frameworks
CMD
DATA
SW
Usage Service Abstraction
CMD
DATA
SW
Users: AF/DoD & Partners
HPMLF Service Abstraction
High Performance Machine Learning Framework (HPMLF)
Parallel ML Processing (batch, interactive, streaming
Scalable & HP Platforms (Parallel/Distributed, GPU Cluster)
Parallel ML Workflow (KestoneML, TensorFlow, SystemML)

Figure 4. High performance machine learning framework (HPMLF).