Open Source Collaborative AI Development in the Enterprise NeuroSystem Group

Dinesh Verma
IBM T J Watson Research Center
1101 Kitchawan Road,
Yorktown Heights, NY 10598, U.S.A.
Phone: +1-914-945-4599
Email: dverma@us.ibm.com

Bill wright
Red Hat
150 Mathilda Place, Suite: 500
Sunnyvale, CA 94086, U.S.A.
Phone: +1-415-710-5207
Email: bwright@redhat.com

John Overton
Kove IO, Inc.
14 N Peoria – 2H
Chicago, IL 60607, U.S.A.
Phone: +1-312-850-3308
Email: john.overton@kove.net

Ravi Sihna
Reliance Jio
15506 Mountain View Lane
Frisco, TX, 75035, U.S.A.
Phone: +1-425-766-7516
Email: ravi.p.sinha@ril.com

Abstract

The Enterprise Neurosystem Group is an AI and machine learning (ML) open source research community effort which includes representatives from several large and small industrial and academic institutions. The goal of the group is to facilitate the development of technologies, documentation and best practices which are needed to create the enterprise nervous system, a large interconnected collection of various AI models that can provide a C-suite executive with insights into all aspects of corporate operations in real-time. The academic partners in this group include Stanford Stanford Linear Accelerator Center, Harvard Analytics and University of California Berkeley's Data-X program. The industrial participants include personnel from America Movil, Reliance Jio, Equinix, Ericsson AI, IBM Research, Intel, Kove, Penguin Computing, PerceptiLabs, Seagate and Yahoo! In this paper, we perform a case study on this group, with an overview of its vision, its structure, and different activities within the group.

Key Words: Open Source, Artificial Intelligence, Community, Case Study

1 Introduction

The many benefits that an enterprise can derive from the use of Artificial Intelligence and Machine Learning technologies are very clear. The growing list of AI-assisted functions include, but are not limited to, network operations and security, supply chains and logistics, autonomous drones (infrastructure inspection), IT operations and security, financial systems and forecasts, natural language-based customer interaction/management, facility and network equipment maintenance, human resources, legal functions, insurance and risk analysis, global and regional regulatory frameworks, taxation analysis, and facilities security/employee verification.

These activities are typically siloed by their function-specific applications from various vendors. This limits the potential for collective inference and ultimately, the business value of AI, as visibility is kept to a small subset of the total data, out of reach of large-scale management and insight.

In addition, differences in vendor AI development philosophies have produced multiple variants of application architectures, and as a result, customers will field AI applications with wildly different degrees of complexity, interpretive capability, and insight. A lack of application-level cohesion contributes to reduced or conflicting bias assessment capabilities. Given the broad experience of data scientists, bias is both unfortunate and a real challenge.

This multi-architecture situation causes fundamental AI operational issues, as a) the various vendors and the underpinnings of their architectures are vastly different, and b) gaining interpretive insights in terms of inference, output, and bias analysis becomes an integration challenge. Different levels of intelligent analysis and self-reflection will lead to conflicting conclusions - just as they do across humanity.

Traditional approaches by enterprises have been to develop their own proprietary solutions to address these issues. However, a new approach to address them using open source cross-organization is being developed as an alternative approach to develop cross-enterprise solutions. Addressing these issues can result in a transformative infrastructure. The Enterprise NeuorSystem Group was formed with a goal of discovering the right way to attain this vision in an open collaborative community.

At the current time, the group consists of volunteers from the following companies and industries: America Movil, Equinix, Ericsson AI, EY, Fiducia AI, Harvard Analytics , IBM, Intel, Kove, Penguin Computing, PerceptiLabs , Red Hat, Reliance Jio, Seagate, Stanford SLAC, UC Berkeley, Verizon Wireless, and Yahoo!

2 Project Genesis

Around the beginning of 2018, some of the founding members of the group, such as Bill Wright of Red Hat, were challenged by a multinational telecommunications provider to examine the role of artificial intelligence in mobile networks. Their goal at that time was to explore the role of AI/ML in these telecommunications networks, at both a granular level and a much larger organizational scale, with operational insights from a leading mobile operator.

As a result of that study, it became clear that while many enterprises are creating AI models for specific analytics functions, those models would tupically run independently from other AI instances. Large-scale correlation of findings, particularly in real-time scenarios, was missing.

This led to the realization that mobile networks and related data centers possess an interesting parallel to the human neurosystem. The human neurosystem is a connective biological framework that links different subsystems and helps each operate autonomously. At the same time, it also correlates different inputs from each of these subsystems when corrective attention is required. It seemed logical to assume that a similar evolutionary framework could emulate this neural topology within the IT domain, and in turn, enable a deeper and more complete level of AI self-awareness and perception. This vision would open the larger possibilities of AI in any enterprise, regardless of the business vertical.

The participants decided to build on the vision for this large-scale collective framework comprising a single AI instance spanning the enterprise. This system would connect the related AI nodes and data sets, bring together findings, and conduct analytics and an analysis of patterns across all company operations. This framework, when employed, would provide the C-suite of an organization with a window into many aspects of its business functions in real-time by identifying trends, predicting challenges, and delivering a choice of optimized solutions. It would include tightly defined, high-value AI functions to drive community interest and user adoption. Concurrently, the group would build a top-tier interpretive and reporting intelligence – one that takes in all forms of real-time and historical data, both structured and unstructured, and responds with either autonomous remedial action, or course-correction recommendations to an organization's management team. It would be assembled and nurtured by a formal community that would welcome operator, enterprise, and vendor participation in an open Research and Development forum. An architecture of this scale and importance should be open source for a variety of reasons: model design and code base transparency, ongoing community input - especially relating to algorithmic bias - and cost efficiency.

The current group was formed as a result of this exploration and grew in size as participants from different companies and academic institutes joined in the vision.

3 Strategic Vision

The group proposes a unified analytics framework that spans all aspects of corporate operations. Much like the neurology of the human body, it will be a series of interconnected AI/ML models, tailored to permeate and assess every aspect of the business – and autonomously regulate and optimize most day-to-day functions. This capability would not only increase efficiencies in every area of IT and business operations, it would also provide a corporate operations interface of unparalleled depth and clarity.

The initial program will start with a small series of open source projects designed to tackle fundamental issues, each with a clear path to value. These will be linked by a common communication and AI/ML framework, which will be developed concurrently. The structure and relationships of these respective efforts will be determined by the members of these projects.

Over time, this connective framework and its associated projects will extend awareness across the enterprise, tying together areas of operation via the web, corporate networks, manufacturing and logistics systems, and IT software systems and databases.

Data integration would then be fine-tuned by corporate users to reframe the deployment and target additional areas of analysis. They would teach the system to understand new areas of focus and further enable the system to be self-directed. Eventually, this AI analytics system would run autonomously by maintaining a communication and analytics web across all aspects of the business.

In its end state, an overarching intelligence would draw in and cross-correlate the data from the many areas of company operations, and autonomously provide load-balancing for lower-level functions - systemically adjusting corporate operations in real time. For larger strategic challenges, human management would be brought in to deliver the best decision capability.

The Enterprise Neurosystem framework should result in improved and more efficient service delivery, deeper customer insights, supply chain and delivery efficiencies, data center operations and environmental savings, facility security optimization, and much more. The resulting savings and increase in revenue would be significant.

To address the creation and deployment of this architecture, high-impactful projects and PoCs will be initiated to build the project's momentum. These initial small steps are each designed to increase operational efficiency in their own right, and in turn, illustrate positive results. Each project will deliver high value and cost savings to the targeted area of focus. The following steps will lay the foundation to arrive at the unified end state. By initially targeting a tight set of high-value deliverables, and sequentially building this framework according to open source principles and best practices, the end result will be a significant advance in F500 business operations.

4 Strategic Architecture

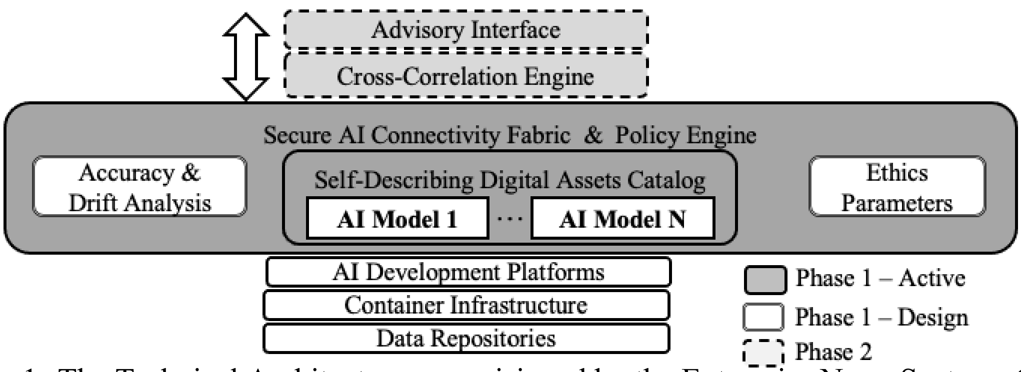

Figure 1: The Technical Architecture as envisioned by the Enterprise NeuroSystems Group

The technical high level architecture of the end-system envisioned by the group members can be seen in Figure 1. The enterprise AI solutions will be built on-top of existing open source packages or products that are available. The ENG assumes the existence of such open source AI packages such as sci-kit learn (Pedregosa et al., 2011) or tensorflow (Abadi, 2016). With modern practices for cloud computing, these would also be implemented using a container framework, and thus be managed by an open source package such as Kubernetes (Brewer, 2015), or equivalent commercial packages such as OpenShift (Picozzi, Hepburn, & O'Connor, 2017). Furthermore, any AI solution has to rely on an assortment of data repositories.

Atop this pedestal of existing software packages, the enterprise neurosystem group envisions three types of capabilities which needs to be built. The first capability is that of an enterprise-wide catalog of models that are developed, used in various places, and reused and built into larger aggregates. These require a catalog framework, and a system for efficient secure interconnection among different parts of the enterprise. These components are being built in the initial phase by the enterprise neurosystems group and are shown as Phase 1 components. Other Phase 1 components which are being discussed include the framework for ethics and an approach to monitor model accuracy and drift.

In the next phase, different AI models needs to be cross-correlated and an interface for the enterprise level advisory functions would be needed. Those are Phase 2 functions which will be worked upon in subsequent stages.

The type of AI models would used in the environment would depend on the use-cases and the enterprise. For managing sustainability, AI models related to climate change, insect ecosystem, or animal habitat may be needed. A telecommunications operator may need AI models for radio access networks, streaming video, network traffic management and customer relationship management functions. A manufacturing plant may need models for supply chain, manufacturing and procurement processes. The strategic architecture is applicable to a large variety of organizations in multiple industry segments.

5 Governance and Working Structure

As the number of participants exceeded over 30 participating organizations, it was felt that a semi-formal organization was needed to govern the set of active volunteers. Towards this objective, an initial charter for the group was developed to govern its operations, and a *Governing Board* for the group was elected using majority voting among all participants. The board would have a life-span of one year, and elections were conducted subject to the constraint that no single organization would have more than two representatives on the governing board. This allowed the governing board to be controlled too strongly by the interests of a single organization and promoted a true open source collaborative experience.

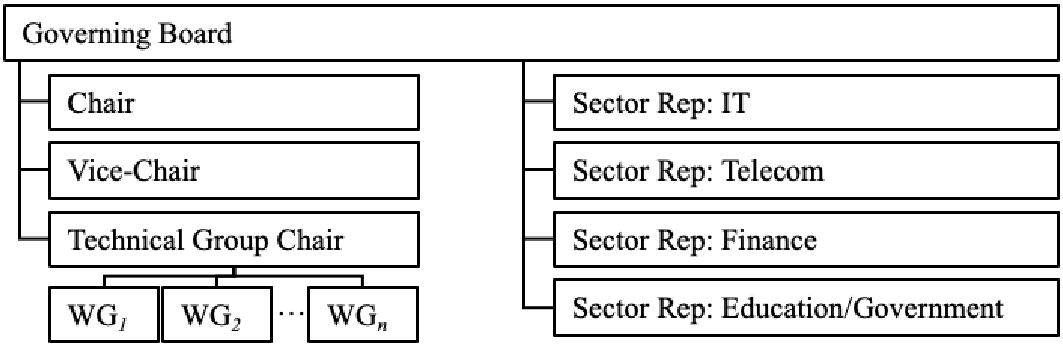

Figure 2: The Governing Board Structure of Enterprise NeuroSystems Group

As per the charter, the governing board consists of three sector-independent leaders and four representatives representing the requirements of four industrial segments as shown in Figure 2. It is responsible for setting the strategic direction of the community and for defining the policies regarding the conduct and all non-technical activities of the group. The majority of the technical work done to develop, advance and promote adoption of the work is expected to be done by community members who are working in different technical working groups.

The sector independent leaders include the Chair of the governing board who is the overall leader of the community, and the Vice-Chair of the governing board who can step in and act on behalf of the Chair when he/she is not available. The third sector independent leader in the group is the technical group leader, who is responsible for the coordination of the different technical activities underway among the different technical working groups. The responsibilities of the Chair, Vice Chair and Technical Group Leader are analogous to the roles of the CEO (Chief Executive Officer), COO (Chief Operating Officer) and CTO (Chief Technical Officer) found in a typical enterprise. In addition to these three sector-independent members of the board, the board consists of technical representatives consisting of experts with background in telecommunications, finance, Information Technology and Government/Education industrial sectors.

The current members of the governing board elected as the first board are as shown in Table 1:

Table 1: Governing Board Members

Board Position	**Name**	**Organization**
Chair	Bill Wright	Red Hat
Vice-Chair	John Overton	Kove
Technical Group Leader	Dinesh Verma	IBM
IT Vendor Rep	Ganesh Harinath	Fiducia AI
Telecom Sector Rep	Tong Zhang	Intel
Finance Sector Rep	Vish Hari	Meta
Education/Finance Sector Rep	Ryan Coffee	Stanford SLAC

6 ENG Technical Organization

The actual work on the group is done by technical working groups, which look at specific topics and develop technologies that leads towards the eventual strategic vision.

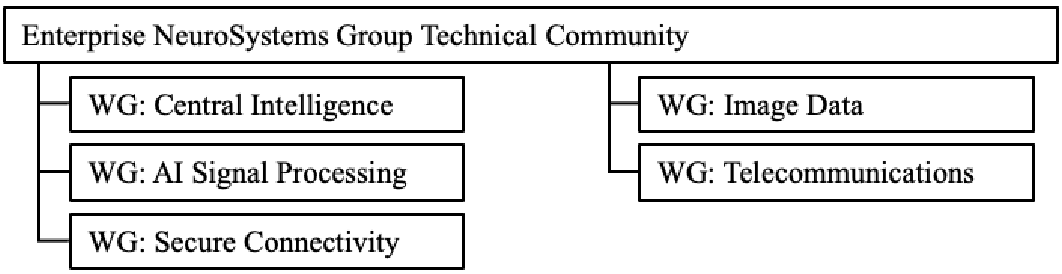

Figure 3: The Technical Structure of Enterprise NeuroSystems Group

Each technical working group is led by a technical volunteer, and consists of participants from two or more participating organizations. The technical groups have their independent tracks of operations, define their scope of activity, and report back periodically to a biweekly meeting of the entire community. At the current time, there are five active working group, the Central Intelligence Working Group, the AI Signal Processing working group, the Secure

Connectivity Working Group, the Imaging Data Generation Working Group and the Telecom Working Group. This technical structure of the group is shown in Figure 3.

This details of each of the working groups and the technologies they are working on are described below.

6.1 Central Intelligence Working Group

The mission of the central intelligence platform is to create a set of common utilities that can be used in other groups and support a large variety of enterprise use-cases. After an intensive set of internal discussions, the group member agreed that the community is lacking an efficient and usable mechanism for self-describing digital assets. Digital assets are data-sets, AI models, rule-sets, reports, programs and similar artifacts that are generated and exchanged during the course of the operation of an AI system. While there are plenty of digital assets that are shared, created, generated and used across different enterprises, digital assets tend to be fragile, exchanged according to preset conventions, and it is hard to determine what an digital asset is from the inspection of its representation alone. This lack of representation causes significant brittleness in the nature of distributed solutions that are developed.

A scheme by which the properties of any digital asset can be determined is needed to address this challenge. If digital assets are self-describing, i.e. they contain their own description, it would be easier to share, search and manage digital assets. However, a good approach for self-description is hard to find given the existence of many legacy digital assets, the need for flexibility so that self-description is applicable to a wide set of use-cases and balancing that against the need to have an efficient approach that is not too onerous for users, or introduces significant inefficiency.

To balance these requirements, a light-weight approach for specifying the metadata for any asset has been developed by this working group. This approach allows for a flexible description of a metadata associated with the digital asset, with specification of a scope for the digital assets and a type within the scope. The scope could identify the publisher or intended using organization of the digital asset and is represented like an organization name, while the type provides for usage within the specific scope. Additional JSON field can provide for additional attributes of the self-describing asset.

The team is creating an initial catalog to store and distribute self-describing assets, along with a few demonstrator use-cases. Concurrently, the team is discussing about potential new common utilities that would be worthwhile developing.

6.2 AI Signal Processing Working Group

AI/ML can be used for processing many types of signals from IoT devices. Signals could be audio/acoustics information from microphones, or they could be vibration data from seismic sensors. Both acoustics and vibration data share many characteristics in that they are time-varying signals of several frequencies superimposed together to create a composite signal. These types of signals can be processed using ML techniques to address a variety of problems such as anomaly detection, classification or root-cause analysis.

A working group is exploring the use of ML on such signals, and their use-cases. The original AI processing software was developed by IBM for acoustic signals and used for building and industrial manufacturing applications. In those environments, sounds can be used to detect how long a machine is operational and schedule maintenance on basis of the runtime. This optimizes the cost of technician dispatch, specially for older equipment which does not have in-built instrumentation. Another application is the detection of sounds in equipment such as squeaks, screeches or grating noises that indicate a problem that requires attention. In manufacturing plants, hissing sounds can be used to check for leaking gases, and a light hit on metal bars can check for possible problems in manufacturing of the rivets. There are multiple applications of acoustic processing. IBM has donated this software to the Enterprise NeuroSystem Group to enable community applications for the larger cause.

One of the activities the Enterprise NeuroSystem Group is conducting in this area is the use of acoustics to monitor bee-hives. Bees play an important part in the ecosystem of the Earth and bees provide pollination for a wide variety of plants, in addition to producing honey. Current techniques to monitor the health of hives are intrusive and can interfere with the operation of the hive. Monitoring the sounds that honeybees produce and using that to analyze the health of the hive provides an alternative approach which is non-intrusive and can be deployed widely.

Creating the technology to inspect and monitor the health of bees in open-source can also help source some of the important ecological challenges facing the world today.

ENG is using the acoustics technology to create solutions for monitoring beehives in a collaborative manner.

6.3 Secure Connectivity Working Group

In order to inter-connect the different digital assets, data sources and AI models that are present within an enterprise, a secure and dynamic infrastructure to connect different disparate systems within an enterprise are needed. In modern cloud based environments, different clusters of a management infrastructure like Kubernetes (Brewer, 2015) are used to implement each individual system. However, connectivity across two such clusters does not always provide security assurances as information is exchanged across clusters. This can cause information leakage unless a network administrator in the enterprise sets up the right network configuration.

An open source project skupper (https://skupper.io) provides an application-level interconnect to enable secure communication across Kubernetes clusters. It uses an application level proxy gateway that can communicate and establish secure communication using transport layer security (TLS) tunnels across different clusters. This open-source provides a convenient and rapid to use secure interconnectivity paradigm.

The working group in ENG is looking at the ways to use skupper to address the enterprise interconnect problems, and produce best practices around it. This working group is an instance of an activity in ENG where the focus is on reusing other open-source components, as opposed to creating a new open source component.

Another aspect being examined is the addition of policies to manage the connectivity among applications. Policy based management (Verma, 2002) is a proven approach for simplifying network management, and the task can be further simplified by using devices that generate their own policies (Calo et al., 2018). This extension is being considered by the working group.

6.4 Imaging Data Generation Working Group

The Immaging data Generation Working Group is led by Stanford SLAC LCLS (Stanford Linear Accelerator Laboratory - Linac Coherent Light Source) group. They are developing a Proof of Concept for the LCLS Cookiebox detector.

The architecture incorporates FFT algorithms to quickly process image data at the edge of the detector, and then forward the data to more powerful NN capability in a core data center. Images currently arrive at a rate of 120 images a second per spectrometer, and this will increase to one million images per second (20M/sec total rate) over the next two years.

AI/ML capability will need to be created to quickly sift through this flood of data in real time, and search for anomalies that indicate new discoveries in physics. This is a concise example of a Neurosystem topology - real-time data processing in multiple nodes, connected to a central cross-correlation intelligence. This can serve as the first step to a broader-based intelligent system, connected to other core data centers for deeper discovery and analysis. The same architecture pattern can be replicated in a wide variety of industries, including telecommunications, manufacturing, financial services and health care.

6.5 Telecom Working Group

Mission of Telecom Working Group is to connect all the dots needed at cognitive wireless connectivity to pass relevant information coming from Management, Control and User payload layers as well as next generation services to the compliant Central intelligence Engine. Federated AI can consume data through the different level of catalogues provided by the Central intelligence framework and further process it for cross correlation with Public, Enterprise as well as Government use cases. Advent of 5G has brought a vital change to the DNA of wireless communication. Introduction of Hardware SW decoupling, A fully distributed 5G architecture, cloud native connectivity platform, intelligent Service Management Orchestration Engine has made 5G a perfect target for Artificial Intelligence based automation and support of Next generation services like Metaverse, Digital twining, Remote Rendering and Mixed reality applications. We envision a world where virtually all devices (eMBB, uRLLC and mMTC) are much more intelligent, simplifying and enhancing our daily lives. 5G subscription is forecasted to reach 4.4 billion in 2027 it becomes universal auto drive connectivity that can efficiently connect virtually everything around us. Telecom Work Group is led by Tier One Operators and vendors in collaboration with academia and research groups and SDOs worldwide. Cognitive RAN, Core and SMO features and functionalities are compliant to ORAN and ONAP specs.

7 Conclusions

In this paper, we have provided an overview of the Enterprise NeuroSystems Group. This group provides an example of a new approach for doing AI/ML in enterprises. We have provided an outline of the strategy and the current tactical execution plan of the group. While the failure or success of this model for developing AI for enterprise will have to be re-examined a few years from now, this provides a new model that has not been exercised before.

References

Abadi, M. (2016). Tensorflow: learning functions at scale. In *Proceedings of the 21st acm sigplan international conference on functional programming* (pp. 1–1).

Brewer, E. A. (2015). Kubernetes and the path to cloud native. In *Proceedings of the sixth acm symposium on cloud computing* (pp. 167–167).

Calo, S., Verma, D., Chakraborty, S., Bertino, E., Lupu, E., & Cirincione, G. (2018). Self-generation of access control policies. In *Proceedings of the 23nd acm on symposium on access control models and technologies* (pp. 39–47).

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … others (2011). Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, *12*, 2825–2830.

Picozzi, S., Hepburn, M., & O'Connor, N. (2017). *Devops with openshift: Cloud deployments made easy*. " O'Reilly Media, Inc.".

Verma, D. C. (2002). Simplifying network administration using policy-based management. *IEEE network*, *16*(2), 20–26.