InnovateOps

David A. Callner

Senior Vice President, Technology
Solutions of RIVA Solutions, Inc.

To keep up with the vast amount of data generated, there must be synergy between the data scientist and operations team. This synergy is created by making quality, transformed data available to the data scientists and the business analysts as soon as it is in the pipeline. This requires end-to-end orchestration and automation of the data pipeline from data ingestion to visualization. RIVA applies DataOps and MLOps best practices to automate a pipeline using our InnovateOps approach. With this approach, the data is immediately available to the data scientists and business analysts for exploration, model building, reporting, and visualization. Multiple ML models can be evaluated, built dynamically upon new data, and pushed to production seamlessly. Later, the models can be monitored automatically through a wide array of visualization tools and techniques.
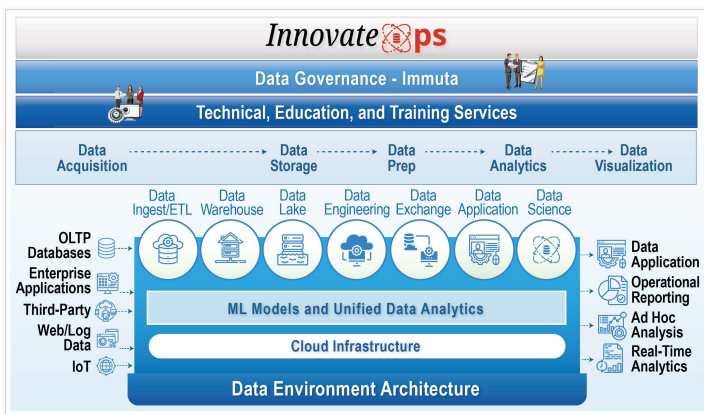
InnovateOps: RIVA works with a wide array of industry-leading technologies across Data Integration, Business Intelligence, ML, and Data Science, enabling our customers to access both enterprise services through an extensive network of connectors, drivers, programming languages, and utilities. We have expertise in many data science products that represent innovative solutions for all aspects of a Data Science pipeline.

Our InnovateOps approach effectively accelerates the processing of data events using practical data science techniques. We address the following challenges using our proactive approaches:

- Focus on supporting full end-to-end data science lifecycles from Data Ingest to Data Frontend Applications
- Allow easy access to data science assets
- Supporting a wide variety of analytic use cases that would use data science
- Techniques and technology that help eliminate manual data wrangling and reduce data wrangling cycle times

- Fusion, integration, and alignment of disparate data sources with a focus on repeatable processes and automation of workflow steps
- Mapping data to the intended use through analytics and visualizations

Data Acquisition: RIVA's InnovateOps is backed by a suite of best-in-class capabilities for data acquisition, preparation, and architecture that run on data and business logic in tandem. Our overall approach includes a combination of vendors and technologies that are adhere our principles and guidelines. We deliver cross functional teams that balance robustness and capability while reducing the number and complexity of components within the Data Acquisition and Preparation phases. Our data acquisition and preparation services enable the import of data for immediate use or further processing via storage in a designated central repository. We use backend data lakes or warehouses via real-time streams or batches for ad hoc or persistent services for on-demand access. We develop architectures that allow for the ingest of both static files and real-time streaming data as it is produced.

Making data visible and accessible starts with data acquisition. RIVA employs a proactive, forward-looking approach to data acquisition focused on anticipating future data requirements. We do not wait in "receive mode" for new data sources and resources to be revealed. RIVA begins by actively engaging stakeholders (such as from data consumer targets and data source owners) through assessments of the current state of the customer's data ecosystem. This will ensure a deep understanding of the current and available data, acquisition workflows, data compliance requirements, and initiatives to acquire and ingest new data. RIVA will then map this ecosystem to identify potential improvements as well as best practices to continue.

Data Prep: RIVA's experience working with AWS, Azure, and Google Cloud offerings benefit our customer by providing them with experts who know how to work with cloud environments to resolve data, ETL, and DataOps/MLOps related issues. We recognize the challenges of working with full datasets. Certain datasets will have millions of rows, and visual ballparking is not sufficient for assessing data quality. To address this issue, RIVA's InnovateOps uses a combination of tools including Kylo, NNDATA, and Trifacta for structuring, cleaning, and aggregating bigger datasets. This allows users to work through the phases of Discovery, Structuring, Cleaning, Enriching and Blending, and Optimizing and Publishing.
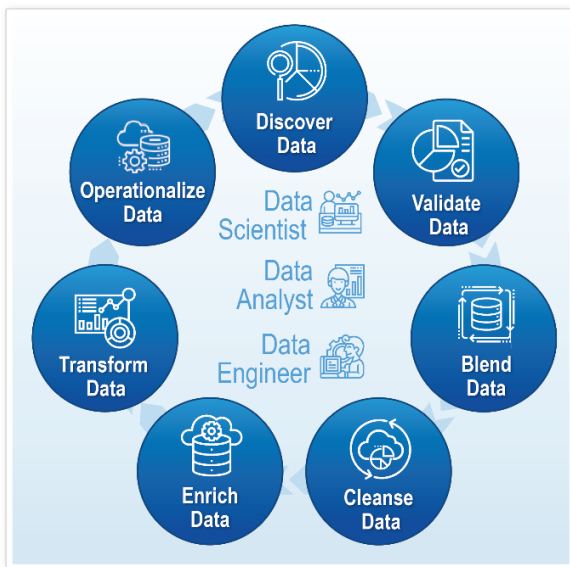
InsightOps Framework

We start with a smaller, more manageable subset of the data to enable fast exploration, iteration, and refinement. From there, a Data Analyst's familiarity with the business questions at hand can accelerate understanding of the dataset and progressively evolve it for the desired end goal. It is essential that the self-service data wrangling software platform is designed for non-technical users specifically, to ease the process of data preparation, including structuring, exploring, and cleaning the data. Three components that are critical to the data preparation process, and that our approach will address are:

Interactive profiling – Allow users to interactively determine the data's shape, size, structure, and outliers.

Predictive transformation –Provide tailored recommendations to perform data cleaning operations to accelerate the data preparation workflow.

Intelligent execution - Compile the user's data cleaning steps and actions to a unified analytics platform (Databricks/Apache Spark) to coordinate jobs across the entire cluster.

In a collaborative environment, it is helpful to have multiple users work on the same assets, or to create copies of quality work to serve as templates for others. A robust data preparation tool (i.e., Trifacta) enables users to collaborate. Our team creates a framework that aligns data preparation flow objects to the data preparation phases. We create and curate a knowledge base of best practices around data auditing, workflow specifications, and workflow execution that can be incorporated as flow objects in a data prep tool. The flow object catalog helps to collate and organize the datasets, recipes, and other objects that we use to generate results more quickly. We collaboratively generate high-quality data pipelines at scale that facilitate use cases, applications, data science, business intelligence, and analytics. Data preparation is a continuous and iterative process, and RIVA has full visibility into the raw data in addition to advanced data

discovery, collaboration, end-to-end data lineage, metadata management, governance, data quality, data security and machine learning-enabled automation that can be seamlessly leveraged across data lakes.

Data Architecture: At the core of our data environment architecture is the ability to quickly spin collaborative workspaces for any data science project. This includes security policies, audits, and identity management, while seamlessly enforcing user and data governance. This enables users to continuously train and deploy ML models for all ML applications, resulting in faster innovation with big data, simplified access to large datasets, and unified data science and data engineering. We will do this by providing:

Collaborative Notebooks – natively support Python, R, SQL, and Scala so practitioners can work together with the languages and libraries of their choice to discover, visualize, and share insights with stakeholders.

Machine Learning Runtime - access to preconfigured ML clusters powered by a scalable and reliable distribution of the most popular ML frameworks and optimizations for unmatched performance at scale

MLOps/DataOps - management of ML models and data from experimentation to production with enterprise security, reliability, and scale

AI/ML/DL Models: RIVA's InnovateOps approach to algorithm development starts with understanding the business/mission problem. Then, through several data exploration and preparation techniques (i.e., variable identification, univariate analysis, bi-variate analysis, and outlier treatment), data is prepared for use by data consumers and data integration across the enterprise. The emergence and maturation of DataOps/MLOps as a practice has enabled analytics, algorithm development, and data processing to be automated in a similar fashion as the software development DevSecOps Agile methodologies. InnovateOps incorporates continuous monitoring and auditing to confirm that the algorithms are accurate. If accuracy decreases, the algorithms are corrected. We provide teams of data scientists who specialize in statistical learning, exploratory data analysis, feature engineering, and algorithm experimentations. These experts are well versed in R, Python, and other languages supported by the enterprise Data Platforms. We focus on algorithm model accuracy, bias, and interpretation that can all deviate as input data changes. Thus, statistical monitoring and standardized validation datasets are provided to track an algorithm's production effectiveness.

The pace of development in data science and AI is rapid, and extensive research is being done on them around the world in academic institutions and by individual data science practitioners. We are dedicated to being on the forefront of these emerging technologies.

This involves constant reviews of open-source forums, working with our academic partners, and participating in conferences and innovation forums such as NVIDIA's GPU Technology Conference (GTC) and AWS re:INVENT to identify strategies and techniques that deliver improved performance over existing products and systems. Knowledge gains are shared with our customers. We share these innovations to advise on the state of advancements in relevant technology solutions to maintain the organization's competitive advantage. In addition, RIVA has fostered technical partnerships with hyper-focused AI/ML companies with practical solutions for applied AI/ML domain that can be applied across a data science workflow. We have developed an inference optimization solution based on research conducted at MIT that can be plugged into a Deep Learning pipeline and integrated into a DLOps optimization process. This solution achieves up to 10x improvement over existing inference engines and provides automated analysis on which AI/ML models can provide the best predictions. These improvements can allow the running of deep learning models to edge devices.

Data Governance: Often data science teams are forced to rely on ad-hoc, complex processes that restrict agility and increased time-to-data or time-to-deploy. A critical element of a sound data governance strategy is a product catalog for searching and exploring data products across the enterprise. A key tenet of InnovateOps is self-service access to data via a programmatic product catalog and that can be accomplished with an automated governance that automatically applies security policies. Our data governance team develops documents like the Entity Relationship Diagrams (ERD), enterprise data dictionary, meta-data repositories, and data environment SOPs.

To handle the increasing number of regulatory and rule-based restrictions to data, system administrators create more roles that represent specific data access combinations. The correct roles must map to data sets. RIVA's InnovateOps integrates governance solutions (such as Immuta) to aid analytics and data science teams quickly move from this state of immature governance to advanced, automated, fine-grained governance using a software based governance layer that acts as a policy enforcement layer between data actors and the data itself. RIVA provides a comprehensive data governance. Our approach preserves the integrity of the data's quality for use at an organizational level, and ensures accountability and auditability.

Data Analytics: RIVA's InnovateOps uses tools such as DataBricks and Snowflake for a variety of data science use cases. For visualization, we support current analytics tools such as Tableau and Databricks, as well as open-source tools like Grafana, D3, Matplotlib, and ggplot which can be integrated into our overall InnovateOps approach to provide visualization and analytics capabilities. RIVA focuses on automated machine learning on multivariate time-series data-sensor data, and mission data to support use cases such as predictive

maintenance for aircraft, vessels and weapons systems, multi-INT fusion, and exploitation in the intelligence community.

Our capabilities are deployed in both cloud and customer air-gapped classified environments. We are focused on ML on multivariate time series data and using that for operational data challenges. The core engine centers around automated feature learning from raw signal data streams using kmeans/random forest analysis approaches while focusing on specific features that are relevant in a signal stream. For time series data we provide efficient detection of relevant anomalies in data. We design data analytic solutions that run with limited guidance (upper and lower bound) where issues manifest themselves (pre-failure modes) and are tuned to detect novel patterns to enable rapid decision making. We deploy using Kubernetes and Docker containers. We support dynamic visualizations exploiting available data sets. We support live stream monitoring capability with the ability to ingest data in live conditions. Once a model is established, the increment can be processed in near real time, allowing faster insights with an end-to-end process.

Furthermore, the visualizations are browser based as well as API enabled so they can be displayed in an enterprise dashboard. Using InnovateOps, we ingest, label, and process nearly any imaging sensor data, supporting automated analysis. Our solutions are mature, convolutional neural networks and comprise all the tools necessary for a computer vision solution. We follow an assisted learning approach to ML that ensures accuracy and performance - not only in data preparation and labelling, but in model development and sustainment.

We empower our customers, regardless of their expertise, with intuitive data analytic capabilities to meet their specific mission needs with professional-quality AI. Tight, iterative loops between labeling and model development modules enable automated labeling, which significantly reduces the time and effort to generate CV training sets on a variety of data including IR full motion video, aerial synthetic aperture radar (SAR), and satellite multi-spectral imagery. We monitor the entire lifecycle, from a development tool for custom computer vision models, as well as a sustainment platform, able to maintain production models at peak performance over their entire lifecycle.

CONCLUSION

Customers are amassing large amounts of time-series data from sensors, electronic signal intelligence, and many other sources. This data largely goes unanalyzed and unexploited due to both its volume and the complexity of analysis. For next-generation operational intelligence centered around exploiting this data to perform predictive maintenance on weapons systems (e.g., aircraft, high-end sensors, facilities and their critical systems) and real-time detection of events of interest (e.g., distinguishing between earthquakes and missile launches on seismometers, anomalous movement patterns of vessels or aircraft), organizations are turning to machine learning. RIVA provides an automated machine learning platform that has proven useful for both predictive maintenance and electronic intelligence exploitation in industries such as automotive, oil and gas, metals and mining, and semiconductors. We have also had early successes in the defense and intelligence sector.

To foster innovation for our customers, RIVA invests in building new capabilities in key technical research and development, bringing the successful outcomes together in an array of integrated emerging technology solutions platforms through our Innovation Lab. It consists of a group of external thought leaders that continually challenges our thinking to solve the government's toughest problems and create potential solutions.

Our unique approach to implementing emerging technologies starts with RIVA's investment in our Innovation Lab, centered around Artificial Intelligence, Agile, Blockchain, DevSecOps, Cybersecurity, Big Data and Data Analytics, and Cloud Solutions to our customers as a corporate arm. RIVA holds monthly meetings with all our Program Managers to evaluate our people, processes, and technologies to identify areas of continuous service improvement. During these meetings, we cross-pollinate services excellence success stories and work together to drive innovation. We evaluate a variety of emerging technologies and how they might benefit each program.

NEXT STEPS

Setting up an automated data pipeline will take time and effort. It involves changing the culture, and how people do business on a day-to-day basis. RIVA's Data Scientist are experienced in educating our teams and government staff to take data and turn it into actionable insights that support our customers' missions.

For more information on this white paper, contact David Callner directly at dcallner@rivasolutionsinc.com.

RIVA

Founded in 2009
Large-small 8(a) SDB with 11 years in GovCon

450+ Employees
RIVA consultants provide customer service at Federal client sites nationwide

23 Locations
Nationwide workforce located across 23 states

8 Prime Schedules/IDIQs
Experienced program, project, and task managers supported by a fully automated system

16 Federal Clients
Portfolio of Federal, Civilian, Defense, Homeland, and Intel

81 Active Contracts
Prime, subcontracts, and task orders supported by over 100+ partnerships

Explosive Growth
#44 on the Inc 5000 fastest growing companies list

Happy Employees
Reputation as Great Place to Work ensures high retention

Low Turnover Rate
90% retention rate on our toughest projects

ABOUT THE AUTHOR

David A. Callner is the Senior Vice President of Technology Solutions for RIVA Solutions, Inc. Mr. Callner is responsible for providing technical capture management of Federal and DoD opportunities to drive business growth. He leads RIVA's corporate technology practices to enable innovative solutions and capabilities and facilitates the implementation of solutions and practices across programs to ensure service delivery excellence.

ABOUT RIVA SOLUTIONS

Headquartered in McLean, Virginia, right outside of our Nation's capital, RIVA Solutions, Inc. (RIVA) is a larger 8(a) small, disadvantaged business (SDB) working with the Federal government to provide innovative best practices in Management Consulting, Science and Engineering, Cyber, Agile, Cloud, IT Operations and Modernization, Artificial Intelligence/ Machine Learning, and Robotic Process Automation.