Data Science Platform: Pentaho & MISTec

Joint White Paper

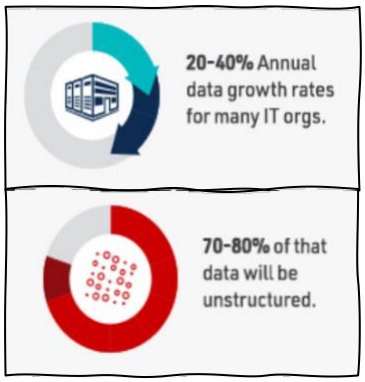
Hitachi Vantara Federal
Design Interactive

Organizations now have more data than ever. Today, the challenge for organizations is how to leverage cutting edge technologies to transform all that data into a real-time intelligence asset for the entire organization.

Data Science Platforms for Digital Transformation

As recent as early 2000s, many organizations viewed data as a liability – content to be archived and kept safe for compliance without any return on investment. With recent advancements in computation, memory, and algorithms, the ability to rapidly convert data to information to aid decision-making processes led to a change in attitude. For the first time, data began to be viewed as an asset and product rather than a liability.

The ability to turn this data into a practical asset is even more important because both public sector and industry are capturing, storing, and analyzing more data than ever before. This presents non-trivial problems to overcome, as the data growth rates are exploding, and much of the new data produced is complex and unstructured. Addressing these criteria and the associated constraints should be a key design component of the next generation Mission Command System and enable the decision dominance at multiple echelon levels based on the widely varying tactical, operational, and strategic needs of a given task.



Even if organizations wish to utilize business intelligence and advanced analytics into their digital strategy, they can still be handicapped by their inability to identify and manage data assets, leverage artificial intelligence (AI) and machine learning (ML), extract real time analytics, and map operational efficiency to optimal solutions.

A successful DataOps-driven data analytics strategy relies on the people, processes, and technologies of an organization. All three must work together in harmony to develop, enable, and maintain a successful and continuously evolving Digital Transformation strategy. A globally connected and information-driven decision-dominant warfighting environment needs this strong data-enriched foundation in order to remain functional, even when the unit is required to operate in a disconnected, denied, intermittent and/or limited (DDIL) environment.

The user experience for decision making and situation awareness should be agile and intuitive to effectively connect the dots and have access to the “single source of truth” while remaining compliant with required security, access controls and regulations. The provenance and validity of the data is prime for data quality assessment and trust in using the same for patterns and trends analytics and the Machine Learning (ML) automation enablement.

The open architecture foundation for a data science platform is key. The critical need for ease integration of data sources is well understood. Irrespective of modality (structured tabular data, character separated values (csv), text, audio, video, image, log extracts, social media, RSS Feeds, emails, application APIs, and more), language, media format and geography. This ease also extends to the integration required with the command and control information environment (C2IE), Continuous Integration/Continuous Deployment (CI/CD) and/or Machine Learning Operations (MLOps and ModelOps).

In short, successfully implementing Digital Transformation strategies rely on four core pillars: **Decision Dominance, Data Governance, Metadata Management, and Data Storage**. This is the non-trivial challenge of Digital Transformation that both Hitachi Vantara Federal (HVF) and Design Interactive (DI) can address effectively with a solution architecture composed of Pentaho DataOps Platform and Military Information Superiority Technology (MISTec) fit to the organizations work environment and data culture.

Hitachi’s Pentaho DataOps Platform

Hitachi Vantara Federal offers an entire suite of data analytics products and services to bring innovation to an organization’s Digital Transformation journey:

- **Pentaho** enables organizations to access, prepare, and analyze all data from any source, in any environment, between data managers and consumers. Self-service analytics is enabled for both business and technical users integrated reporting and dashboard needs. R and Python script executors enable Machine Learning (ML)-based object detection modeling and its use as part of the data processing pipeline as well as analyzing the identified object statistics and enabling triggers/alerts based on anomaly detection or other specific customer requirements.
- **Pentaho Data Catalog** is a data self-service platform. It utilizes machine learning to perform automatic discovery and tagging of data, including semantic inference, and can refine tags based off analyst feedback then use the same feedback to refine future tagging suggestions.

- **Hitachi Content Intelligence (HCI)** can tie into existing structured and unstructured data sources and perform transformations to enhance the searchability and compliance of the data.
- **Hitachi Content Software for File (HCSF)** provides fast distributed file system for high-performance computing, AI/ML, and analytics workloads.

Note: *These products are technically mature and immediately demonstrable*.

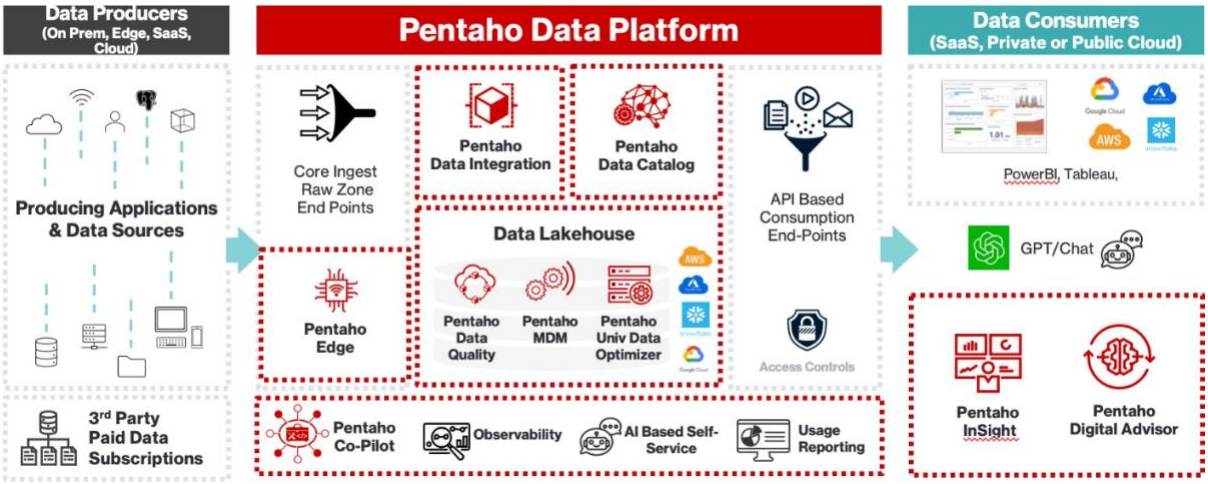


Figure 1 – Pentaho DataOps Suite Features End-to-End Data Governance

This architecture allows data scientists, analysts, and users to reduce the time it takes to process captured data by accelerating metadata discovery and categorizing data. It permits data curators and stewards to manage and control access to sensitive data. It additionally enables IT administrators to eliminate unintended redundant data and configure data tiering (including dynamic and virtualized).

Design Interactive's MISTec Framework

MISTec uses a customized, scalable Artificial Intelligence (AI)-driven algorithm to monitor, evaluate, and prioritize the mass of relevant information from disparate sources so Soldiers can observe, orient, decide, and act faster than the adversary despite the reduction in physical resources and computing power associated with mobile command posts (CPs) of the future. The MISTec concept and prototype integrates data analytics and human performance engineering methods to enable Information Superiority, defined as the operational advantage derived from the ability to collect, process, and disseminate an uninterrupted flow of information. Designed with a flexible architecture, the system's components (an event prioritization algorithm, event queuing algorithm, and simulation testbed) support easy modification of prioritization factors, preferences, roles, and commander-critical information requirements (CCIRs). This enables the solution to be scalable with new and evolving CP systems. This design also encourages the maturation of the algorithm components over future research phases, leading to more sophisticated information prioritization.

The MISTec technical solution is composed of three parts: the event prioritization algorithm, the event queuing algorithm, and the simulation testbed. The queuing algorithm relies on the priorities created by the prioritization algorithm and is responsible for managing how events move through the network, considering networking constraints, user rules, and cognitive capacities. The simulation testbed includes synthetic data for validating and testing the prioritization and queuing algorithm.

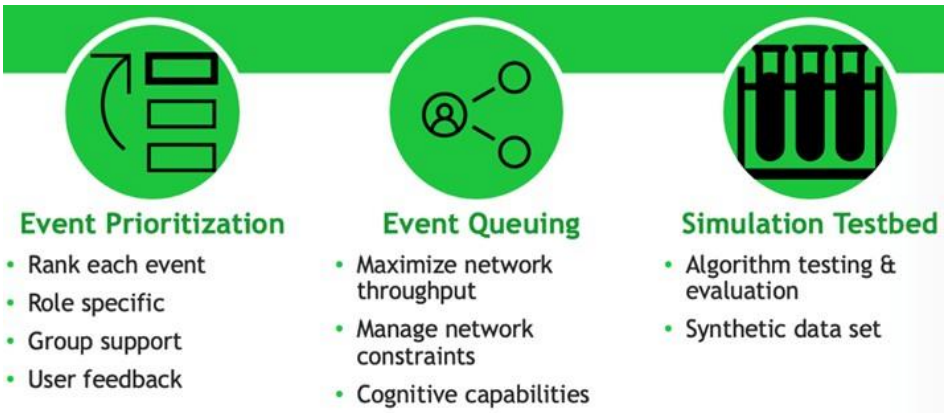
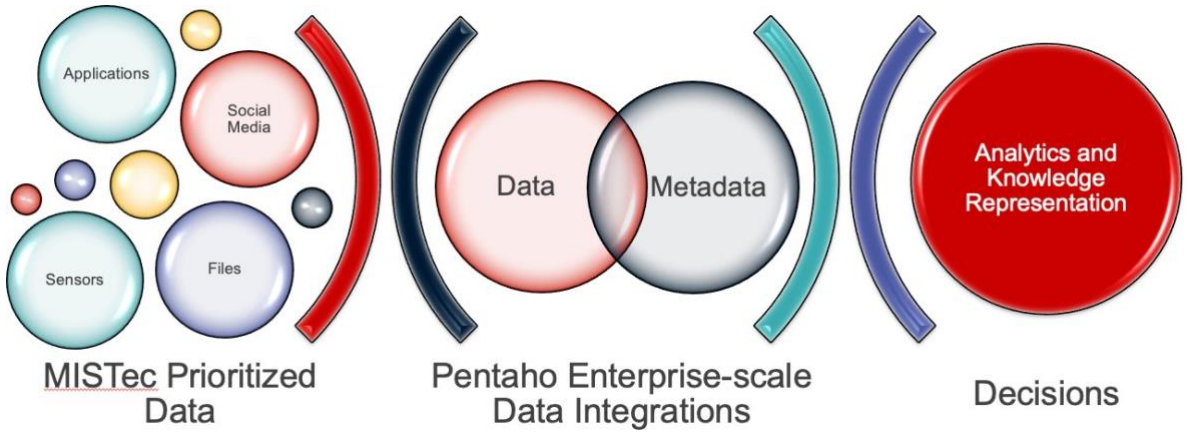


Figure 2 – MISTec is composed of three algorithms responsible for prioritization, message delivery, and testing.

Commanders implicitly think of their staff as information managers and rely on them to filter large amounts of information related to CCIRs to support decision making. The MISTec prioritization algorithm partially addresses this need by relating events to dynamic CCIRs to support decision making under uncertainty and changing conditions.

MISTec provides prioritized information to the right Soldier at the right time based on the commander's guidance in a rapidly evolving situation. This is essential to supporting Mobile Command Posts (MCPs) that are constantly inundated with data that needs to be prioritized to support decision making cycles. Utilizing MISTec will assist with supporting smaller and more mobile, dynamic command posts. This helps with outpacing the enemy's decision cycles by utilizing advanced technology to intelligently manage information. Combining Pentaho's data management capabilities with MISTec helps increase survivability by incorporating the latest hardware and methods while reducing they systems' footprint.



About Hitachi Vantara Federal

Hitachi Vantara Federal is the most reliable data solutions provider with the most complete set of hybrid cloud and data fabric solutions in the Federal IT market. We're a collaborative, full-service company empowering data-driven insight with a deep bench of integrated partners advancing Federal customer missions regardless of their data maturity levels.

We're more than just a leader in digital infrastructure. We're constantly evolving our data management, analytics, and Internet of Things (IoT) solutions to service Federal leaders' mission needs today, and in the future. All of our Federal customers get the platinum treatment as a standard offering. From civilian agencies to the Department of Defense (DoD) and Intelligence Community (IC), we're known for our agility, dependability, and excellence.

As the trusted leader in mission-centric data solutions for the Federal government, Hitachi Vantara Federal is dedicated to providing the solutions and support our customers need today, and in the future. We power digital transformation from the core to the edge, offering a full-service experience for integrated digital solutions spanning the entire data fabric. Balancing a century of tradition with constant innovation, we stand for agility, dependability, and excellence–guaranteed. We have a TS facility clearance and we are familiar and adhere to the NIST AI Risk Management Framework, DoD Ethical Principles for Artificial Intelligence and the fast-emerging Responsible and Trustworthy AI Guidelines from various Federal agencies as part of our solution implementations. Our analytics products are part of the DoD CDAO Tradewinds Marketplace as well as part of DIA Needipedia (ID# = 06APR2023_01MAR02023_HITACHI-VANTARA_6-1-3_ENG_SUM).

Get in Touch

WEB: hitachivantarafederal.com/about/contact-us/
EMAIL: info@hitachivantarafederal.com
PHONE: 1-703-787-2900

About Design Interactive

Design Interactive, Inc. (DI) is a woman owned small business that applies a human factors-centric approach to the research and development of technologies and methodologies to accelerate mastery and optimize human performance. Since 1998 DI has been providing engineering services and consulting in human-system integration (HSI), training systems design and evaluation, next-generation HSI research, and usability services. As the company has matured, our customer base has continued to grow, including increasing work from commercial entertainment, manufacturing, and automotive industries.

Get in Touch

WEB: designinteractive.net/contact-us/
PHONE: 1-407-706-0977