Redbooks
ibm.com/redbooks

IBM Power Systems
Enterprise AI Solutions

Glen Corneau

Andrew Laidlaw

Marcos Quezada

Power Systems

IBM®

Redpaper

IBM Redbooks

IBM Power Systems Enterprise AI Solutions

September 2019

REDP-5556-00

Note: Before using this information and the product it supports, read the information in “Notices” on page v.

First Edition (September 2019)

This edition applies to the following products:

- IBM Watson Machine Learning Accelerator Version 1 Release 2 Modification 1 (product number 5765-AEI)
- IBM PowerAI Vision Version 1 Release 1 Modification 4 (product number 5765-TAI)
- IBM Watson Machine Learning Community Edition Version 1 Release 6 Modification 1 (product number 5765-PAI)
- IBM Watson Studio Local Version 2 Release 0 (product number 5737-D37)
- IBM Video Analytics Version 1 Release 0 (product number 5737-L13)
- IBM Spectrum Scale Version 5 Release 0 Modification 3 (product number 5765-F34)
- IBM Spectrum Discover Version 2 Release 0 (product number 5737-I32)
- IBM Watson OpenScale Version 2, Release 1 (product number 5737-K82).

© Copyright International Business Machines Corporation 2019. All rights reserved.
Note to U.S. Government Users Restricted Rights -- Use, duplication or disclosure restricted by GSA ADP Schedule Contract with IBM Corp.

Contents

© Copyright IBM Corp. 2019. All rights reserved.

Notices

This information was developed for products and services offered in the US. This material might be available from IBM in other languages. However, you may be required to own a copy of the product or product version in that language in order to access it.

IBM may not offer the products, services, or features discussed in this document in other countries. Consult your local IBM representative for information on the products and services currently available in your area. Any reference to an IBM product, program, or service is not intended to state or imply that only that IBM product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any IBM intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any non-IBM product, program, or service.

IBM may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not grant you any license to these patents. You can send license inquiries, in writing, to:
IBM Director of Licensing, IBM Corporation, North Castle Drive, MD-NC119, Armonk, NY 10504-1785, US

INTERNATIONAL BUSINESS MACHINES CORPORATION PROVIDES THIS PUBLICATION "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some jurisdictions do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. IBM may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

Any references in this information to non-IBM websites are provided for convenience only and do not in any manner serve as an endorsement of those websites. The materials at those websites are not part of the materials for this IBM product and use of those websites is at your own risk.

IBM may use or distribute any of the information you provide in any way it believes appropriate without incurring any obligation to you.

The performance data and client examples cited are presented for illustrative purposes only. Actual performance results may vary depending on specific configurations and operating conditions.

Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

Statements regarding IBM's future direction or intent are subject to change or withdrawal without notice, and represent goals and objectives only.

This information contains examples of data and reports used in daily business operations. To illustrate them as completely as possible, the examples include the names of individuals, companies, brands, and products. All of these names are fictitious and any similarity to actual people or business enterprises is entirely coincidental.

COPYRIGHT LICENSE:

This information contains sample application programs in source language, which illustrate programming techniques on various operating platforms. You may copy, modify, and distribute these sample programs in any form without payment to IBM, for the purposes of developing, using, marketing or distributing application programs conforming to the application programming interface for the operating platform for which the sample programs are written. These examples have not been thoroughly tested under all conditions. IBM, therefore, cannot guarantee or imply reliability, serviceability, or function of these programs. The sample programs are provided "AS IS", without warranty of any kind. IBM shall not be liable for any damages arising out of your use of the sample programs.

© Copyright IBM Corp. 2019. All rights reserved.

Trademarks

IBM, the IBM logo, and ibm.com are trademarks or registered trademarks of International Business Machines Corporation, registered in many jurisdictions worldwide. Other product and service names might be trademarks of IBM or other companies. A current list of IBM trademarks is available on the web at “Copyright and trademark information” at http://www.ibm.com/legal/copytrade.shtml

The following terms are trademarks or registered trademarks of International Business Machines Corporation, and might also be trademarks or registered trademarks in other countries.

AIX®
IBM®
IBM Cloud™
IBM Elastic Storage®
IBM FlashSystem®
IBM Research™
IBM Spectrum®
IBM Spectrum Conductor®
IBM Watson®
POWER9™
Redbooks®
Redbooks (logo) ®
Watson™

The following terms are trademarks of other companies:

Linux is a trademark of Linus Torvalds in the United States, other countries, or both.

Java, and all Java-based trademarks and logos are trademarks or registered trademarks of Oracle and/or its affiliates.

RStudio, and the RStudio logo are registered trademarks of RStudio, Inc.

Other company, product, or service names may be trademarks or service marks of others.

Preface

This IBM® Redpaper publication helps the line of business (LOB), data science, and information technology (IT) teams develop an information architecture (IA) for their enterprise artificial intelligence (AI) environment. It describes the challenges that are faced by the three roles when creating and deploying enterprise AI solutions, and how they can collaborate for best results.

This publication also highlights the capabilities of the IBM Cognitive Systems and AI solutions:

- IBM Watson® Machine Learning Community Edition
- IBM Watson Machine Learning Accelerator (WMLA)
- IBM PowerAI Vision
- IBM Watson Machine Learning
- IBM Watson Studio Local
- IBM Video Analytics
- H2O Driverless AI
- IBM Spectrum® Scale
- IBM Spectrum Discover

This publication examines the challenges through five different use case examples:

- Artificial vision
- Natural language processing (NLP)
- Planning for the future
- Machine learning (ML)
- AI teaming and collaboration

This publication targets readers from LOBs, data science teams, and IT departments, and anyone that is interested in understanding how to build an IA to support enterprise AI development and deployment.

Authors

This paper was produced by a team of specialists from around the world working at IBM Redbooks, Austin Center.

Glen Corneau is a Partner Client Technical Specialist for Cognitive Systems in the United States of America. He has 28 years of experience in IBM Power Systems running IBM AIX® and Linux. He holds a bachelor's degree in computer science from Texas A&M University at College Station, Texas. His areas of expertise include high-performance computing, systems management, cloud, and cognitive computing. He has presented extensively at IBM Power Systems Technical Conferences and is an IBM Redbooks® author.

Andrew Laidlaw is the senior AI infrastructure specialist in the UK. He is a Level 2 Certified Technical Specialist with 6 years of experience in IBM Systems, during which time he has worked with the latest technologies and developments. His areas of expertise include open source technologies that include Linux, open source databases, and AI frameworks and tools. He has presented extensively on these topics worldwide, including at the IBM Systems Technical University conferences. He has been an author of multiple previous Redbooks publications.

© Copyright IBM Corp. 2019. All rights reserved.

Marcos Quezada is a Fulbright Scholar with a master's degree in management information systems from Northern Illinois University. He concentrated his studies in business analytics. Since 2017, he has the role of Cognitive Systems Technical Leader for Spanish South America. He has 20 years of experience in the IT sector. He consults on enterprise AI, ML, deep learning (DL), modern data platforms, cloud computing, and other open source technologies. He holds a degree in systems engineering from Universidad de Belgrano in Argentina. He has been an author of multiple previous Redbooks publications.

The project that produced this publication was managed by:
Scott Vetter, PMP

Thanks to the following people for their contributions to this project:

Ann Lund
IBM Redbooks

Freddy Alves Vaquero, Luis Armenta, Ivaylo Bozhinov, Scott Campbell, Srinivas Chitiveli, Tom Farrand, Eric Fiala, Nitin Kapoor, Bo Ran Lee, Sam Matzek, Gustavo Santos, Scott Soutter
IBM

Now you can become a published author, too!

Here's an opportunity to spotlight your skills, grow your career, and become a published author—all at the same time! Join an IBM Redbooks residency project and help write a book in your area of expertise, while honing your experience using leading-edge technologies. Your efforts will help to increase product acceptance and customer satisfaction, as you expand your network of technical contacts and relationships. Residencies run from two to six weeks in length, and you can participate either in person or as a remote resident working from your home base.

Find out more about the residency program, browse the residency index, and apply online at:

ibm.com/redbooks/residencies.html

Comments welcome

Your comments are important to us!

We want our papers to be as helpful as possible. Send us your comments about this paper or other IBM Redbooks publications in one of the following ways:

- Use the online **Contact us** review Redbooks form found at:

  ibm.com/redbooks

- Send your comments in an email to:

  redbooks@us.ibm.com

- Mail your comments to:

  IBM Corporation, IBM Redbooks
  Dept. HYTD Mail Station P099
  2455 South Road
  Poughkeepsie, NY 12601-5400

Stay connected to IBM Redbooks

- Find us on Facebook:

  http://www.facebook.com/IBMRedbooks

- Follow us on Twitter:

  http://twitter.com/ibmredbooks

- Look for us on LinkedIn:

  http://www.linkedin.com/groups?home=&gid=2130806

- Explore new Redbooks publications, residencies, and workshops with the IBM Redbooks weekly newsletter:

  https://www.redbooks.ibm.com/Redbooks.nsf/subscribe?OpenForm

- Stay current on recent Redbooks publications with RSS Feeds:

  http://www.redbooks.ibm.com/rss.html

1

Introduction

This publication’s goal is to promote the understanding among three roles: line of business (LOB); data science; and information technology (IT) professionals, around the definition of an organization’s information architecture (IA) for enterprise artificial intelligence (AI) solutions. This IA must support the development stages for enterprise AI solutions while addressing the needs of the enterprise AI workflow.

It is time for enterprise AI solutions to go from the proof of concept (PoC) stage to production at scale. This task can be achieved only when these three roles have a common understanding about what is needed from an IA to accelerate the enterprise AI workflow.

We show how IBM Cognitive Systems and AI offerings and capabilities can address these needs at the intersection of each role and our five selected use cases so that your business is ready to understand your organization’s first enterprise AI business scenario, problem statement, and use case.

© Copyright IBM Corp. 2019. All rights reserved.

1.1 Enterprise AI

Enterprise AI differs from consumer AI because it needs an infrastructure that can adapt to changing business needs at scale. It can start as small as a PoC running on a marketing manager's laptop to a global solution with AI models being generated in a central location and pushed out for federated inferencing on edge devices.

All these tasks happen on a hybrid infrastructure supporting an IA that can feed data effectively into the four stages of the AI workflow while keeping the enterprise AI solution available and secure.

IBM Cognitive Systems provide an industry-leading enterprise AI infrastructure for machine learning (ML), deep learning (DL), and inference to fuel new thinking and capabilities across your business. Your organization needs AI infused into every business process where heuristics are used today to drive greater confidence in business decisions at scale.

An enterprise AI solution must be implemented on an infrastructure that supports specific AI project needs and complies with traditional mission critical demands and requirements:

- Faster time to results and accuracy
- Increased resource utilization
- Simplified management
- Enterprise-grade solution

By meeting these demands and requirements with an enterprise AI infrastructure, your AI solutions can grow with your organization to make people's jobs easier and more productive and to make the best use of processes and resources by expanding open source-based AI innovation in the enterprise.

1.1.1 AI workflow

Throughout this publication, we describe the enterprise AI workflow. This section introduces the four main stages: ingest, preparation, training, and inference, and it mentions some of the challenges that are faced at each of the stages.

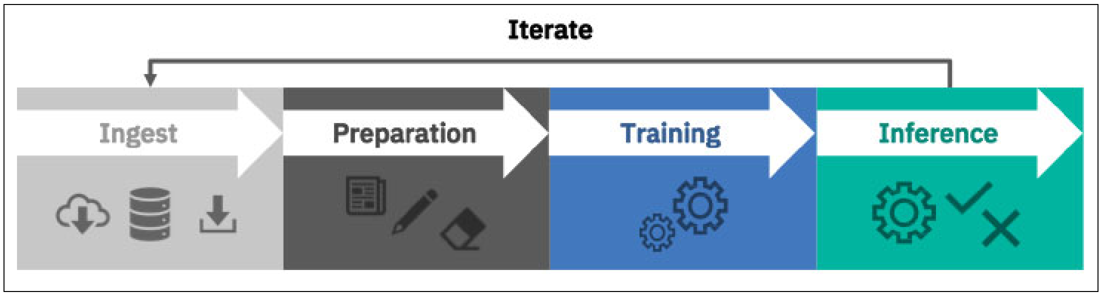


Figure 1-1 The AI workflow

Other sources of literature on the subject might have greater or fewer stages. These stages are expansions of substages or compressed stages of the workflow that we use in this publication. A four stage workflow is used to illustrate the examples that we present here.

Ingest

This stage is about bringing the main resource into the AI workflow: data. Data may come from many sources and in many formats. Speed here is key, and data movement must be minimized to feed the subsequent stages.

Challenges of this stage include:

- Multiple sources and formats of data
- Multiple types of data structure: structured, semi-structured, and unstructured data
- Large volumes of data to be managed
- Unknown, unorganized, and unlabeled data

Preparation

The preparation stage is where approximately 80% of the AI workflow time is spent. At this stage, it is critical to allow the data science team to be effective at their incumbent tasks while simultaneously working with other members of the organization like subject matter experts (SMEs). SMEs can more effectively label data because they understand the domain of interest for which a model is being built while data scientists can focus on other tasks:

- Data visualization
- Data cleaning
- Data validation for completeness or for bias
- Data augmentation
- Data splitting
- Data delivery

Challenges at this stage include:

- Raw data conversion
- Non-existent metadata
- Non-standard data
- Unlabeled data

Training

After the data set is ready, the important work of training is undertaken by the data science team. This stage is compute-intensive because data scientists try many different combinations of variables and algorithms to get to the best model with the highest accuracy. Their tasks include:

- Model selection
- Model preparation
- Model tuning
- Variable (or feature) engineering
- Hyperparameter tuning
- Iterative training runs

Challenges at this stage include:

- Increased resource demand when many AI projects hit the training stage.
- Training scaling up and down over multiple compute nodes.
- Reducing data movement to bring data in and out of training cycles.
- Data lifecycle management of training data sets.

Inference

After the training phase delivers the best model for the business challenge, put it to work. The inference phase includes:

- Model delivery
- Model implementation
- Deploying the model for operational use

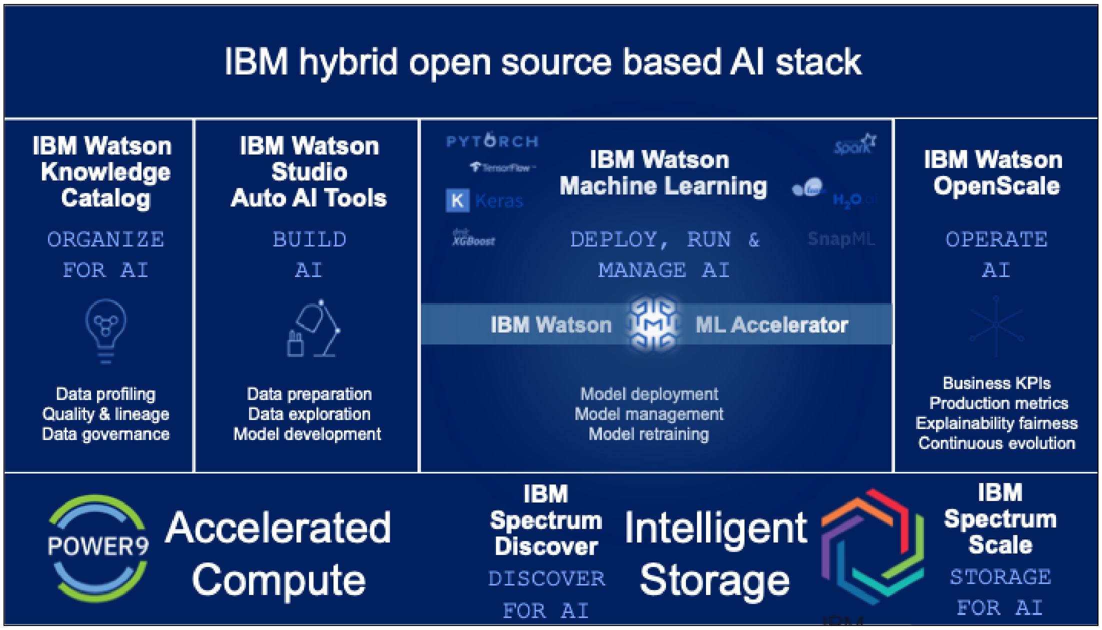
Challenges at this stage include:

- Deploying and managing multiple models, versions, and data pipelines.
- Constantly checking your model for unwanted bias.
- Inferencing scaling up and down across multiple compute nodes.
- Regularly checking model performance and maintaining accuracy.

1.2 IBM Cognitive Systems offerings

This section introduces the IBM Cognitive Systems offerings. It includes a summary of the offerings that are available and their capabilities and where to go for more information.

IBM AI offerings deliver a truly hybrid, open source-based AI stack that is positioned to support your AI implementations from the POC stages to production at scale, as shown in Figure 1-2.


IBM hybrid open source based AI stack

IBM Watson Knowledge Catalog
ORGANIZE FOR AI
Data profiling
Quality & lineage
Data governance

IBM Watson Studio Auto AI Tools
BUILD AI
Data preparation
Data exploration
Model development

IBM Watson Machine Learning
DEPLOY, RUN & MANAGE AI
IBM Watson ML Accelerator
Model deployment
Model management
Model retraining

IBM Watson OpenScale
OPERATE AI
Business KPIs
Production metrics
Explainability fairness
Continuous evolution

POWER9 Accelerated Compute

IBM Spectrum Discover
DISCOVER FOR AI

Intelligent Storage

IBM Spectrum Scale
STORAGE FOR AI

Figure 1-2 IBM hybrid open source-based AI stack

1.2.1 AI cognitive infrastructure

Here are the IBM Power Systems and IBM Storage hardware offerings that are designed to support your on-premises AI workflow:

- IBM Power Systems AC922
- IBM Power Systems LC921
- IBM Power Systems LC922
- IBM Elastic Storage® Server
- IBM FlashSystem® storage family

For more information about the hardware offerings to support your AI use, see the following publications.

- *Cognitive Computing Featuring the IBM Power System AC922*, REDP-5555
- *IBM PowerAI: Deep Learning Unleashed on IBM Power Systems Servers*, SG24-8409

- *IBM Power System AC922 Technical Overview and Introduction*, REDP-5494
- *IBM Power Systems LC921 and LC922: Technical Overview and Introduction*, REDP-5495
- *Introduction Guide to the IBM Elastic Storage Server*, REDP-5253
- *IBM FlashSystem 9100 Product Guide*, REDP-5524

1.2.2 Organizing for AI

These offerings provide the means to organize, store, and process your data to make it readily available for your AI projects.

Data volume and value continue to grow. Performance and access are critical to realizing the fastest time to insight. If your data is not available to your business quickly, then opportunities might be lost. The best way for the IT department to be aligned to their organization's AI business needs is to ensure fast and available data access with an IA that is tailored to that need.

Before AI spread across every industry sector, data governance was considered something to be tolerated to pass an audit. In the AI era, data governance is a competitive advantage for understanding where data comes from; what it is about; and who is entitled to use it throughout the entire AI workflow.

IBM Spectrum Scale

IBM Spectrum Scale is a unified file and object software-defined storage (SDS) for high-performance and large-scale workloads for hybrid cloud. It offers automated, policy-driven, tiered storage management that matches the cost of storage to the value of the data being stored. Integrated information lifecycle tools can manage exabytes of data and billions of files, providing rapid access regardless of how it is tiered.

Coupled with a Hadoop implementation, IBM Spectrum Scale can replace Hadoop Distributed File System (HDFS) to reduce both storage requirements up to three times and accelerate parallel processing up to three times as well.

Other capabilities include:

- Advanced file management, routing, and caching capabilities
- Data dispersal and erasure coding for faster rebuild
- Data protection through snapshots, replication, and backup
- Data encryption and governance for compliance
- Cryptographically secure erasure of sensitive data

For more information, see IBM Spectrum Scale.

IBM Spectrum Discover

IBM Spectrum Discover is a modern metadata management software that provides data insight for exabyte-scale unstructured storage. IBM Spectrum Discover easily connects to multiple file and object storage systems both on-premises and in the cloud to rapidly ingest, consolidate, and index metadata for billions of files and objects, providing a rich metadata layer on top of these storage sources. This metadata enables the data science team to efficiently manage, classify, and gain insights from massive amounts of unstructured data.

Other capabilities include:

- Automated cataloging of unstructured data by capturing metadata as it is created.
- Support multiple file and object storage systems from IBM and other vendors.
- Enable comprehensive insight by combining system metadata with custom tags to increase storage administration and data consumer productivity.
- Leverage extensibility by using the Action Agent API, custom tags, and policy-based workflows to orchestrate deeper content inspection and organize data to support the AI workflow.

For more information, see IBM Spectrum Discover.

IBM Watson Knowledge Catalog

As IBM Spectrum Discover tackles the challenge of data governance and data readiness for unstructured data, IBM Watson Knowledge Catalog deals with the same challenges of structured data.

Other capabilities include:

- Discover more relevant assets more quickly. Interactively discover, cleanse, and prepare your data with a built-in data refinery.
- Curate and shape analytical assets, including data, ML models, and notebooks.
- Protect data misuse and confidently share assets with automated and dynamic masking of sensitive data elements, and govern with active policy management.
- Understand data quality, data lineage and distribution through data-profile visualizations, and built-in charts and statistics.

For more information, see IBM Watson Knowledge Catalog.

1.2.3 Building AI

After your labeled data sets are ready, start working on getting the best model with the highest accuracy. This task varies as data science professionals try various methods. It is a good approach to include tools that give shared access to non data-science professionals, who have knowledge of the meaningful business questions in the industry sector.

IBM Watson Machine Learning Community Edition

IBM Watson Machine Learning Community Edition (Watson ML CE) can get your data science team members set up and operating as quickly as possible. It is delivered as a set of software packages that can deploy an ML environment within minutes by using a few simple commands.

The software distributions are pre-compiled and include everything that you need to build and manage a distributed environment, including the DL frameworks and any supporting software components that they require to run.

For more information, see IBM Cognitive Systems developer portal.

IBM Watson Studio

IBM Watson Studio is a data science and ML platform. It helps enterprises simplify the path from PoC to production at scale; speeds data exploration, and model development and training; and helps scale data science operations across the AI workflow.

With IBM Watson Studio, organizations can tap into data assets and inject predictions into business processes and modern applications. It is suitable for hybrid multicloud environments that demand mission-critical performance, security, and governance in public clouds, private clouds, on-premises, and on the desktop.

Other capabilities include:

- Automatic data preparation, feature engineering, hyperparameter optimization, and ensembling.
- Explore data and use ML with enhanced visual modeling.
- Visually program for DL with an intuitive drag, no code interface in Neural Network Modeler.

For more information, see IBM Watson Studio.

IBM PowerAI Vision

IBM PowerAI Vision automates the DL workflow for visual data like images and video. IBM PowerAI Vision provides tools and interfaces for business analysts, SMEs, and developers without any skills in DL technologies to begin using DL techniques. This enterprise-grade software provides a complete infrastructure to label raw data sets for training, creating, and deploying DL models. It can help train highly accurate models for classification and object detection use cases.

Other capabilities include:

- Rapidly identify and label data sets.
- Train and validate a model in a GUI interface.
- Streamline model training.
- Use existing models as a starting point for faster time to accuracy.
- Deploy an API with one click based on a trained model to integrate into applications.
- Manage both raw and labeled data.
- Video object detection and labeling assistance.

Videos that you import can be scanned for objects, and the objects can be automatically labeled.

For more information, see IBM Cognitive Systems developer portal.

H2O Driverless AI

Driverless AI from H2O.ai employs techniques from expert data scientists in an easy to use application to scale your data science efforts and push your AI initiatives to the finish line of the AI workflow. Every data science professional, domain scientist, and LOB can develop trusted ML models. This automatic ML platform includes functions for data visualization, feature engineering, model interpretability, and low-latency deployment.

Other capabilities include:

- Automatic feature engineering
- ML interpretability
- Natural language processing (NLP)
- Automatic scoring pipelines

- Support for time series data
- Bring Your own Recipes (BYOR) capability

For more information, see H2O Driverless AI.

1.2.4 Deploying, running, and managing AI

After you achieved the accuracy that is required for your trained model, put it to work. With IBM Cognitive Systems and AI solutions, you can deploy models on the cloud or on-premises with a hybrid cloud infrastructure.

An enterprise AI solution must be cost-effective, efficient, and secure. It must also rely on an infrastructure that can adapt to changing business demands and is also tailored to support specific data science needs for automation, data science productivity tools, and security.

The need to increase resource utilization is a key objective for many enterprise clients that know exactly how to accomplish this task for their point of sale systems or their data warehouses, but they might struggle to do this task with their ML and DL tasks.

IBM Watson Machine Learning Accelerator

IBM Watson Machine Learning Accelerator (WMLA) supports the complete AI workflow. It includes open source AI frameworks in an integrated and supported package with enhancements to help you address large and complex enterprise AI solution implementations.

Capabilities include:

- The most popular frameworks (TensorFlow, PyTorch, and Keras) with the ability to use your own framework.
- Bigger training jobs or shorter completion times by distribution across multiple GPUs and nodes.
- Achieve better accuracy and insights without compromise and support for large models.
- Faster time to results and accuracy reduces the time for any single operation within the AI workflow, which can result in hours, days, or weeks of saved processing time.
- Increased resource utilization by scaling model training and inferencing to get the work done in a reasonable amount of time with more accuracy for data models, and real time centralized or on the edge inferencing.
- Simplified management of AI tasks. Everything from ingestion and preparation of data to classifying data, feature engineering, hyperparameter tuning, and simplified deployment of models.
- Enterprise solution capabilities with the highest levels of security, availability, scalability, reliability, and support from IBM:
  - Authentication: Production support for Kerberos, SiteMinder, Active Directory and LDAP, and OS authentication.
  - Authorization: Fine-grained access control (role-based access control (RBAC)), Spark version lifecycle management, notebook updates, deployments, resource plans, reporting, monitoring, log retrieval, and execution.
  - Impersonation: Different LOBs can define production execution users.
  - Encryption: SSL and daemon authentication. Storage encryption with IBM Spectrum Scale.

For more information, see IBM Cognitive Systems developer portal.

1.2.5 Operating AI

After your model is in production, you must understand how it is performing and when it must go through the enterprise AI workflow again to maintain accuracy. Depending on the use case, it might need to integrate with other tools to build a solution. An audit might require you to provide explanations for your model's results. For these requirements, your IA must bring trust and transparency to the operation of your models.

IBM Watson OpenScale

The IBM Watson OpenScale platform tracks and measures outcomes from enterprise AI solutions across its lifecycle, and it adapts and governs enterprise AI to changing business situations.

Other capabilities include:

- Track model performance of production AI and its impact on business goals, with actionable metrics in a single console.
- Apply business results to create a continuous feedback loop that improves and sustains enterprise AI outcomes.
- Maintain regulatory compliance by tracking and explaining AI decisions across workflows, and intelligently detect and correct unwanted bias to improve outcomes.

For more information, see IBM Watson OpenScale.

IBM Video Analytics

IBM Video Analytics turns digital video into valuable information. IBM Video Analytics generates a rich set of metadata that describes moving objects in a stream of video or a recorded video. The video metadata that IBM Video Analytics generates is indexed and stored in a database for future reference, supporting rapid searches, correlation, and analysis.

Other capabilities include:

- Monitoring and identifying events of interest.
- Forensic searching on recorded video footage.
- Find events, objects, and people of interest that appear in recorded video.
- Statistical analysis to identify patterns across hundreds of millions of past events and activities.
- Retrieve video from a mobile video camera, and then search for events, people, and specific objects of interest.

For more information, see IBM Knowledge Center.

1.3 How this paper is organized

This paper's goal is to promote understanding between LOB, data science, and IT professionals. AI solutions are ready to go from PoC to production at scale, which can be achieved only by having tight communication between these three teams to define and create an IA that can support their mixed needs and accelerate the AI workflow.

To promote this understanding we focus on three professional personas or roles:

Line of business These professionals know the business needs and have the most valuable asset when it comes to create and support the business case behind an enterprise AI solution: “The business questions.”

Data science These professionals are the glue that binds the business questions with the ML and DL techniques to create the models fueling the enterprise AI solutions that provide answers.

Information technology These professionals are responsible for the creation of the IA supporting the most valuable resource when building and enterprise AI solution: “Data.”

We look at the intersection of their needs by presenting five categories of use case:

- Artificial vision
- NLP
- Planning for the future
- ML
- AI teaming and collaboration

These use cases were selected by the authors from the ones that generated the most interest around the world. We then brainstormed and came up with 50 questions or concerns that were asked in hundreds of client engagements and grouped their answers at the relevant intersections of a persona point of view and a use case.

With this creation process, we created a flexible publication that can be read in its entirety, with a focus on a persona or on a particular use case across the personas.

1.3.1 Selected AI use cases

Although there are a wide range of applications for cognitive technologies, we are using examples from selected use cases for this publication. Within these general categories, there are various specific examples of where the technologies are used across multiple industries.

Artificial vision

Artificial vision can be used for business challenges involving large amounts of visual data, like images or video streams:

- Identifying abandoned bags at an airport by using existing security camera feeds.
- Finding faults in processor chips by using high definition images from the production line.
- Automatically finding boundaries of properties for insurance valuations.

Natural language processing

NLP uses techniques and technology so that you can interact more naturally with computer systems:

- Classify document types and pull out relevant information to complete a process in a bank.
- Identify caller sentiment and inform a customer service agent in a telecommunication’s call center.
- Automatically transcribe calls to make the content searchable for law enforcement.

Planning for the future

AI solutions can be used to provide more accurate predictions for future outcomes based on historical data, which can help inform decision-making processes:

- Use historical sales records to predict future demand based on weather and events.
- Replace at-risk parts on a drilling machine proactively before a failure stops work.
- Predict and prevent fraudulent credit card transactions as they happen.

Machine learning

With ML techniques, businesses can find the trends and patterns within the data they hold in both structured and unstructured forms:

- Identify anomalies from standard patterns in network traffic to spot intrusions.
- Automate the processing of loan applications based on historical approvals.
- Stream sensor data from IoT devices to identify faults in solar panel arrays.

AI teaming and collaboration

The best results for cognitive solutions generally arise when teams can share their data and expertise and collaborate to create an enterprise AI solution:

- Data scientists collaborate with SMEs to build specific models.
- Common data sets like weather information or footfall data are shared between teams.
- Multiple data scientists collaborate on a single project to pool their expertise.

Throughout this publication, we explore the details of our offerings in relation to these use case categories. However, the capabilities apply across a wider range of applications than we cover here. The specific business challenge that you are working to solve influences the most appropriate tools to use.

2

Point of view: Line of business

Information technology (IT) has spread to all processes within organizations. Conversely, in recent years the IT area has changed from being consulted on every technology decision to being the department responsible for the operation of existing IT infrastructure. As almost all business initiatives have a strong technology component, line of business (LOB) professionals have taken ownership of IT investment budgets, IT project steering committee decisions, and even the definition of what technology solutions to implement. AI solutions are no exception to this trend, and we argue that it is where the previously described phenomenon is strongest.

LOB professionals are concerned with finding answers to business questions, and those answers should be supported by evidence. With AI, they can discover that evidence based on data and not static rules. This data is produced by their own organizations and the environments in which they operate. For those answers to have business value, they must be presented to the correct people at the correct time. LOB professionals should pay close attention to their organization's information infrastructure strategy to ensure that it supports the discovery of evidence at the same pace as data is produced by their company and its surrounding environment.

AI can help LOBs to answer the following kinds of business questions:

- How much or how many? Sales that are forecast per region.
- Which category? Is this a positive or a negative blog post about our product?
- What group does this fall into? Customer segmentation.
- Is this non-standard or a questionable transaction? Anomaly or fraud detection.
- What options should we take? Next best action.

To sum up: LOB professionals want AI to help them guide their business choices, deepen customer engagement, and capitalize on new sources of revenue by predicting the future and embedding AI into business processes to get value from the valuable company data generated and stored over the years. This data includes both traditional structured data, such as relational databases, and unstructured data like text documents, images, sound, or geographic coordinates.

In this section we go through five use cases to understand how business needs can be better addressed, from the proof of concept (PoC) to the production stage, by creating solutions based on IBM Cognitive Systems AI offerings.

© Copyright IBM Corp. 2019. All rights reserved.

2.1 Artificial vision

One of the most common use cases for deep learning (DL) technology in business today is in the field of artificial vision, also known as computer vision. With these techniques, businesses use large quantities of visual data, including images and video, to train a DL model to:

- Classify images into categories.
- Identify objects within images or video.
- Find an object's boundaries within images or video.
- Identify specific actions or movements within video.

Using these capabilities, businesses create a range of models to help them process visual data faster and more efficiently than human-centric processes do. The models that are created can be general-purpose, such as identifying people within a video, or more bespoke for a business and their visual data, which solve specific problems.

Real world example: A utilities company uses drones to inspect their power lines, and has integrated artificial vision to identify faults in need of repair in real time. The system can then automatically order new parts and create work orders. This system has reduced inspection and reporting time by 90%, and increased the number of inspections by 10x.

Using tools like IBM PowerAI Vision, businesses can build DL models that use their own data and their own expertise. With these models, they can build a range of enterprise AI solutions for the challenges they face, reducing expenditure on visual inspection, reacting more rapidly to specific situations, or creating innovative capabilities to enhance their business or offer new products or services.

Some examples of the use of artificial vision include:

- Automated systems to recognize and react to people in controlled areas.
- Identify faults or imperfections at every station on a production line to reduce waste.
- Detect aggressive or inappropriate activity and alert the appropriate authorities.
- Count the number of infected cells in a blood sample or on a pathology slide.
- Real-time translation of sign language to improve communication.

To create a high-quality DL model for artificial vision, you need a labeled data set of example images or video. The quality of this labeling is crucial because this is the data that is used to train the system. Quantity is also an important factor because the more varied the labeled examples are, the higher your likelihood of getting a high-value model.

IBM PowerAI Vision includes labeling tools that subject matter experts (SMEs) can use to quickly and clearly label the images in a data set. With the interface, they can categorize images and label objects or actions of interest without needing to work with a data science team or learn specialized tools. There are more functions that are available to increase the speed of creating a large and high-quality data set:

Auto labeling Labeling new images by using a model that is trained on a smaller number of images, which then requires confirmation of the new labels rather than manually adding them.

Data augmentation Creating images by taking images that are already labeled and performing transformations, like rotation or blurring.

Using these techniques, a SME can rapidly label a data set and train a DL model to help solve a real business challenge. This model can then be deployed as an API service, and the model can be used as needed or integrated into a specific application.

Artificial vision models that are trained by using IBM PowerAI Vision can also be used to identify objects in live or recorded video streams by using IBM Video Analytics. This solution works with various video sources and automatically identifies and tracks objects of interest, and trigger alerts or responses as needed. By using this combination of technologies, it is possible to track objects of specific interest to your business in live video. There are also a number of pre-trained models for common uses, such as identifying people or vehicles, and categorizing by color. These models make it possible to easily search video information for items of interest.

Real world example: Artificial vision was used in pathology to aid the diagnosis of certain types of cancer by acting as a second reader to confirm the diagnosis of a trained pathologist. Samples from patients were uploaded to IBM PowerAI Vision, and then labeled by a team of trained pathologists to produce a data set. A model was trained, and then deployed to support new diagnoses. Artificial vision reduces the time of a confirmed diagnosis from 2 weeks to 2 hours because there is no need to wait for a second opinion from a different pathologist.

By using artificial vision, businesses can create capabilities based on image and video data, which can help improve performance, reduce costs, or develop new revenue streams.

2.2 Natural language processing

Natural language processing (NLP) is the set of technologies and methodologies from linguistics, computer science, information engineering, and AI that deals with how computers process and analyze natural language data.

Business solutions based on these technologies and methodologies include:

Document classifiers Understanding document content based on the text content alone for classification and processing of those documents within business workflow.

Cognitive optical character recognition (OCR) Understanding document content within specific graphic formats for processes based on the content's information and location within those documents.

Voice-of-customer Understanding the sentiment and entity relationships behind a tweet, blog post, or email to spot trends in customer feedback, identify business opportunities, address concerns, reduce churn, and drive revenue.

Ad optimization Understanding relationships between entities, recognizing named entities, and using sentiment analysis for effective placement of advertisements.

Virtual assistants Understanding what a client is asking through a customer relationship channel.

Real world example: Cognitive OCR is useful when information value comes from the data and from where that data is located. Some countries or states require that preliminary voting results are published within an hour of closing polling stations. Forms containing the count from voting boxes at a polling location are electronically sent to a central location where an AI-based system then validates each polling station's official stamps and signatures, candidate names, and the vote tally.

Many NLP solutions can be created by using APIs like Watson Natural Language Understanding, which is available on IBM Cloud™. But, there are use case requirements and industry standards that might require an on-premises infrastructure:

- Intellectual property control.
- Privacy or confidentiality concerns.
- Storage volumes are required.
- Processing (inference or training) time or latency.

IBM Cognitive Systems can comply with all the previous requirements and adapt to the skill levels within an organization when it comes to creating NLP business solutions.

Most major advances in NLP come from research and academia and use open source technology. IBM Cognitive Systems private cloud infrastructure and tools are based on open source standards so that your enterprise NLP solutions can take advantage of the latest developments of NLP techniques.

If there are limited data science skills within your organization, you still can answer business questions by using NLP solutions. The fastest path to results starts with an auto AI toolset. These tools should tackle several challenges in an AI solution workflow besides the skill gap:

- Support the development workflow of an AI solution. Many organizations do not know what steps or stages it takes to build these solutions, so an auto AI tool can help. It is common in the development process to tailor the stages to your company's needs.
- Provide descriptive visualizations and insights to understand the data.
- Find, build, and extract better features. Features are also called *variables*. Variables are the inputs the AI models use to learn and then predict based on new data.
- Build better models fast.
- Infuse trust and transparency by using machine learning interpretability (MLI).
- Facilitate and accelerate how these models are put into production.

If your solution deals only with understanding text, then H2O Driverless AI is the best tool for auto machine learning (ML). When you must understand where that text is located in the document, use IBM PowerAI Vision.

Your organization might have all the required data science skills. In this situation, it is important to start small with quick wins for the business. You can achieve this task by using IBM Watson Machine Learning Community Edition (Watson ML CE). Just as users do not download the Linux kernel source plus all the associated GNU utilities and compile them into a distribution that works for their business, downloading and compiling all the required frameworks for ML and DL is not efficient. Watson ML CE is built mostly on open source software and supported by IBM. It includes compiled and optimized versions of the most commonly used AI frameworks and libraries. IBM also extends these open source tools so they can take advantage of enterprise GPU-accelerated infrastructure. This software stack will have your AI development teams building AI solutions in no time.

When a PoC is ready for production, the next steps involve model lifecycle management and deployment. For that task, you can use Watson Machine Learning. When multiple teams start developing multiple ML or DL models, you must build your company's AI cluster with IBM Watson Machine Learning Accelerator (WMLA).

All of these options need access to the same resource: data. The cloud is great for uploading your data, but quickly stops being financially viable when your solution must bring that data back to your local offices, for example. There are more challenges when you upload confidential data to the cloud. Many vendors talk about the speed with which they can transfer data from their storage subsystems to where it needs to be processed. When dealing with NLP, there is another important speed: the speed with which you identify and select the correct data to build the data sets for model training. IBM Spectrum Discover helps you automate metadata creation and enable rapid metadata querying to accelerate the speed with which you find and collate the set of documents that your data scientists require.

2.3 Planning for the future

From a LOB perspective, incorporating artificial intelligence (AI) to help guide business choices based on intelligent predictive modeling is a key goal. Embedding predictive analytics by using AI into business processes to give more accurate results relies on using data that the business has stored over a period of years.

Incorporating ML into analytics and decision-making can be used across many industries. For example:

- Insurance: Detect fraudulent claims and optimize insurance quotes.
- Commercial banking: Score credit risk and perform fraud detection for credit charges.
- Energy/Utilities: Forecast production and demand and predict outages.
- Government: Predict fraud, and optimize waste systems and traffic.
- Manufacturing: Model product quality and defect detection, and optimize logistics.
- Retail: Customer loyalty, cross-sell/up-sell, and accurate demand forecasting.
- Healthcare: Medical research, and predict patient condition change.
- Transportation: Optimize route planning, and do predictive maintenance.

Real world example: A bank's data scientists were challenged to expand the credit card services of existing customers, easily determine credit risks with better accuracy, and better predict payment defaults. Implementing AI on IBM Power Systems with H2O Driverless AI led to an increased number of credit products that are accepted per customer with more accurate cross-selling for increased revenue.

IBM has set up a partnership with a leading automated ML vendor, H2O.ai, to bring their H2O Driverless AI ("a Data Scientist in a box") product to the IBM GPU-accelerated IBM Power Systems AC922 server.

2.4 Machine learning

In recent years, the use of ML technologies has grown exponentially. As businesses collect more data, traditional analytics techniques no longer scale effectively to provide insight. With ML techniques, businesses can find patterns within the data that they hold or can access so that they can make decisions based on these patterns. Some examples of the use of ML include:

- Predicting future sales from historic sales data, weather information, and other sources.
- Automating common processes to reduce the time that is spent on manual input.
- Identifying anomalies among data, such as security breaches or faults in a system.
- Predicting fraud within financial information and identifying risks.

By using ML techniques, businesses can better use the data that they collect to inform decision-making processes, reduce expenditure, or discover new sources of income. These decisions are supported by data, and similar processes can be used to monitor their effectiveness over time.

Real world example: An enterprise content management solutions provider has embedded ML models and capabilities into their offering to help customers manage the vast quantities of unstructured data they are collecting. They have added functions that are based on ML, including tools to automatically process sales orders, classify new data coming in, and search for potential General Data Protection Regulation (GDPR) violations.

To create a successful ML project, you must understand the business challenges that you are trying to solve, and then have the correct data and skills in place to find the required insights. The time that is taken to build up an appropriate solution for your challenge is affected by many factors, including:

- The quality and quantity of the data sources in use, and how well they fit together.
- Using the right algorithms and techniques on that data, which is an iterative process.
- Performance of the training platform that is used to train and test the models that are created.
- How effectively a team of people can work together to share knowledge and experience.

IBM Cognitive Systems offers a range of tools and technologies to help build ML solutions for bespoke use cases. They can be used to accelerate the time to business value in various ways to get value from an ML project faster.

Watson ML CE is an easy to install software bundle that includes a number of common ML and DL frameworks and tools that are optimized for Power Systems hardware. Within this bundle are the Snap ML libraries, which provide CPU- and GPU-accelerated versions of common ML algorithms. With these libraries, data scientists can train their models up to 46x faster than on other systems, which means that they iterate more rapidly and get higher accuracy in less time. These capabilities are directly compatible with the most commonly used ML libraries, so data science professionals can get the benefits with minimal changes to their existing code.

Another way to increase performance and efficiency is to create a scalable shared environment for your organization’s data science professionals to use. With WMLA, you can create an environment for teams to collaborate on ML projects, and use the resources that they need in a flexible and scalable way. This service ensures efficiency because data and resources can be shared among them, along with the insight and experience brought by the individuals. Watson Studio extends this capability further by adding data management, curation, and sharing across multiple projects and teams to speed innovation.

In businesses where ML skills are not as prevalent, then toolsets like H2O Driverless AI from H2O.ai provide automated ML capabilities. Within the GUI, it is possible to quickly visualize the data that is available, and then create models by using multiple ML algorithms to determine the best option for the data. This approach speeds the time to get a high accuracy model so that data scientists spend more time identifying and integrating appropriate data sources.

Real world example: A bank is using H2O Driverless AI to build models that identify the credit risk of their customers more accurately and faster than previous systems so that the bank can handle 90% of credit applications through these systems, offering more credit products while better predicting payment defaults of existing customers.

ML can transform the way that businesses operate by making data-driven decisions and identifying ways to reduce costs or increase income. Data science teams can use the correct tools to share data and content, train models faster, and run more iterations to get higher business value.

2.5 AI teaming and collaboration

Businesses today implementing a unified AI strategy are discovering the need to balance the data science requirements for computing resources; realistic limitations of actual physical resources (on or off-premises); and the business priorities. Providing a service where the users of AI resources can work together effectively requires both thoughtful planning and flexible implementation, but offers benefits over users competing for resources.

Competing requirements come from all parts of the business:

- Data science professionals want enough GPU-accelerated computing capacity to build (train) models to a level of accuracy that meets the business requirements.
- Data science professionals also want an environment that has the tools they know rather than having to learn skills, with cloud-like rapid provisioning.
- The IT group wants to provide an environment that is secure, highly available, and scalable to meet the business demands.
- The LOBs and the company executives want all the above at a price that meets budgets and allows prioritization based on the business needs.

Real world example: A collaboration between two universities (a medical center and an imaging science institute) used IBM Power Systems servers and WMLA to increase the precision with which doctors were able to identify cancerous brain tumors in under-sampled MRI images. The DL MRI image model was trained in 20x less time versus the x86 server previously used.

GPU resources are always too few and in great demand, so the ability to intelligently use those resources can be paramount. WMLA provides advanced GPU scheduling with an enterprise orchestrator that implements the business priorities (including preemption) across a cluster of servers:

- Support for multiple-GPU and multiple-node training can decrease time to results.
- Dynamic GPU allocation and de-allocation for training jobs by using the Elastic Distributed Training (EDT) capability.
- Full multi-tenancy with customizable role-based access to the managed cluster of servers, which can include both existing x86 GPU-enabled servers in addition to the IBM Power Systems servers.

For collaborative model design, IBM Watson Studio provides a simple web-based portal for data exploration, data preparation, and model development. Popular notebook technologies that re used by data science professionals are included and operate in a private or hybrid cloud environment. The generated models can then be run on the processing resource that makes more sense for the business.

For artificial vision requirements where still images or video are processed through a DL model, IBM PowerAI Vision allows the SME from the LOB to quickly create image classification, object detection, or action detection models by using a simple GUI with no coding required. This tool helps the organization to democratize access to this technology and reduce dependency on the data science team and IT teams.

Data is at the heart of every company and critical for all enterprise AI solutions. Having a common repository available for your GPU-accelerated servers enables many of the advanced capabilities that are mentioned above. IBM Spectrum Scale is a high-performance, highly available, and scalable file management software-defined storage (SDS) solution that supports access to petabytes of data under a single namespace, which reduces the need for multiple copies. Distributed DL requires high-speed access to shared data, which IBM Spectrum Scale provides either in a pre-configured building block (storage and management servers) or as software only, running on existing storage hardware.

3

Point of view: Data science

In today's businesses, the job of the data science team is ever more important. The significance of data and data science increases as businesses look to reduce costs, automate processes, or create innovative products or services. Often, the data science team is called upon to bridge departments and link the business requirements of a project with the information technology (IT) capabilities that are required to solve the challenge. They might be required to understand:

- The business challenge being solved.
- How and where to access the correct data sources.
- What the data availability means in real terms.
- Programming skills, particularly Python or R.
- Machine learning (ML) and deep learning (DL) techniques.
- IT service management.
- Resource management and optimization.
- How to monitor and verify the results of a project.

The artificial intelligence (AI) offerings on IBM Power Systems servers provide capabilities to reduce the burden on data engineers, data scientists, and data analysts when working on ML and DL enterprise AI solutions. With these offerings, the data science team can concentrate on higher-value tasks.

Common areas of focus for tools and functions

There are some common areas of focus among the tools and functions that are available to the data science team, which are aimed at helping you become more productive.

Ease of use

Provide tools that simplify or automate the common tasks that are performed, including:

- Easy discovery and curation of appropriate business data from multiple sources.
- Tools that subject matter experts (SMEs) can use to label data in a quick and efficient way.
- Capabilities to rapidly iterate over multiple data sets and algorithms to discover the best combinations.

© Copyright IBM Corp. 2019. All rights reserved.

- Automate common tasks and repetitive actions like data visualization and hyperparameter tuning.
- Deploy trained models for testing and inference in automated and repeatable ways.

Collaboration

Work together with your team and line of business (LOB) to share expertise and experiences. Use tools and functions so that people can work together more easily:

- Work together on data sets, notebooks, or whole projects with ease.
- Collaborate with SMEs to curate and prepare the best data for a project.
- Share data sets and example code to help tackle new challenges.

Increasing performance

Create models with higher levels of accuracy faster, allowing for a more iterative approach to a project, and allowing your team to test out more algorithms and parameter sets:

- Get faster results by using the appropriate accelerators and techniques.
- Run multiple variations of a DL or ML training run concurrently to compare performance, and then optimize.
- Distribute training runs across multiple accelerators or systems to increase the performance and reduce the time to value.
- Optimize hyperparameters when training models through automatic evaluation and recommendations.

Improving efficiency

Get the most out of your limited resources by using the correct tools to manage, monitor, and optimize the working environment. Improve overall performance by increasing the usage of the resources available to the team.

- Schedule workloads effectively across all of your systems in a flexible manner.
- Work on multiple projects in a single environment without risking data governance issues.
- Deploy new environments rapidly to allow for new projects to be created or ideas to be tested.

Depending on the ML or DL projects that you are undertaking, different tools are available to cover these focus areas. There are some tools like IBM PowerAI Vision that are targeted at image data and building artificial vision models, and others like IBM Watson Machine Learning Accelerator (WMLA) support multiple data types and collaborative projects. The choice of the most appropriate tools depends on the type of project being undertaken.

In this section, we go through our five use cases to understand how the data science teams within a business can benefit from having the correct tools and capabilities at their disposal. They can improve efficiency, collaborate more effectively, and build data-driven solutions more rapidly by creating enterprise AI solutions that are based on IBM Cognitive Systems artificial intelligence (AI) offerings.

3.1 Artificial vision

There is a growing demand for artificial vision solutions in many businesses. As the quantity of visual data including images and videos increases, the need to automate the processing of this data and the responses to this information increases. This situation puts pressure on data engineers and data scientists who are required to develop and train neural networks to classify images or detect specific objects. In many of the use cases, the objects of interest are bespoke industry objects, often with subtle differences between categories. So, standard products can be unsuitable because they are trained for another business' needs.

To create a high accuracy DL model for artificial vision, you need:

- A suitably large data set of relevant images and video.
- Ground truth labeling of the data to match the business need.
- A defined DL network definition for visual data.
- Suitable computational power to process the training iterations.

To create a large set of labeled data, most data science professionals must work closely with SMEs in the industry to ensure that the images and video available are labeled correctly. This process can be time-consuming if the data set is large enough to build a highly accurate model. This process includes the collection and curation of image data along with data normalization and the actual labeling of each example. There is also the challenge of collating the data set in the first place because few businesses have a rigorous storage and organization policy for image data, and so it is often dispersed across multiple systems in various formats.

Using IBM Spectrum Discover, data can be searched across multiple systems and services within a business, searching by file type, metadata entry, or custom tags. This capability makes it possible to pull out all of the image files within a business that are labelled as a certain type, or were created by a particular imaging system. With this tool, you can automatically add custom tags to files as they are created or processed, for example, allowing you to track project usage for governance.

Labeling of images can also be simplified by using the tools that are available within IBM PowerAI Vision. This interface can be used by SMEs so that they can label images and video for classification, object detection, object segmentation, and action recognition. By passing the labeling tasks to people who truly understand the content of the data, it relieves some of the pressure on the data science team to concentrate on building the right neural networks and training and testing models.

You also can use IBM PowerAI Vision to rapidly train models with subsets of the data set, which can then be used to label further images to add to those data sets. The SMEs then simply need to confirm or repair the labels that are generated, speeding up the process (autolabeling). IBM PowerAI Vision has built-in features to quickly increase the size of a data set by adding new labeled examples from existing ones through transformations like rotate, blur, and colorize (data augmentation).

After a data set is curated and labeled, you can use the tools in IBM PowerAI Vision to quickly train a model for that data by using neural network definitions that are designed for:

- Image classification
- Object detection
- Object segmentation
- Action recognition

You can also add your own custom network designs so that SMEs can train and retrain their own models based on your network designs as they adjust their data sets to increase accuracy or eliminate bias. Alternatively, the labeled data sets can be exported in open formats, allowing them to be used in other data science tools. Watson ML CE includes the most common DL frameworks, which can be accessed by using tools like Jupyter Notebooks or data science workbenches like Watson Studio to build your own neural networks and train models.

It is also important to train and test the generated model regularly so that any biases or misrepresentations in the data set can be fixed either by adding more data examples or correcting any mislabeled images. Iterative training helps build a more accurate DL model for your use case, with regular feedback about the performance of the model. Waiting for models to train before selecting the next step can be tedious, so reduced training times are beneficial.

For DL workloads like artificial vision, the use of GPU accelerators increases performance over CPU-only training. GPU processors are designed to run multiple simple calculations in parallel at large scale, and so are ideally suited for the multiple iterations over data sets that DL training requires. The wide range of AI offerings on Power Systems servers are all supported on GPU-accelerated servers like the IBM Power Systems AC922 server. This server supports NVIDIA Tesla V100 GPUs connected to the IBM POWER9™ CPU by using NVLink 2.0 connectivity, offering larger data bandwidth to the GPUs to maximize their performance. These benefits can be seen in the base DL frameworks from Watson ML CE and the higher-level tools like IBM PowerAI Vision. With these benefits, you can reduce training times and iterate more rapidly to help tune your data set, hyperparameters, and model to get the best value.

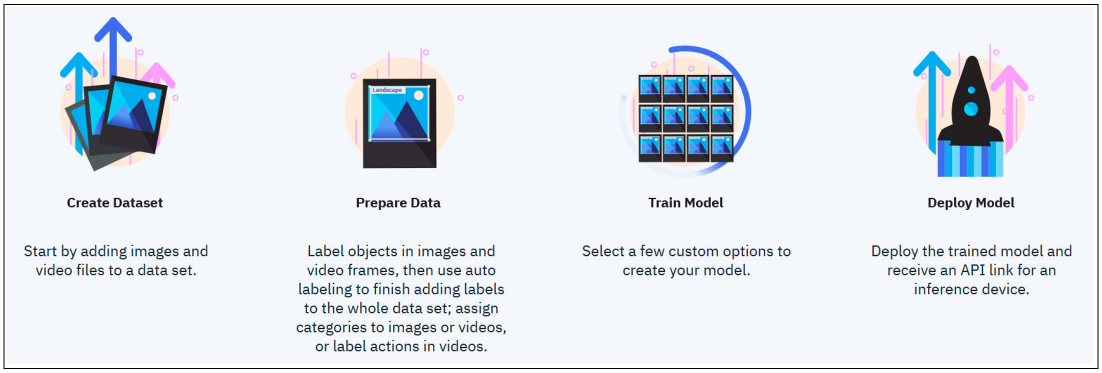
Another way to maximize the performance of DL training is to spread the work across multiple GPU devices or even multiple systems within a cluster. Watson ML CE includes Distributed Deep Learning (DDL), which data scientists can use to distribute a training run across multiple GPUs and multiple nodes by automating the transfer of data, network definition, and parameters among systems to share the processing and produce faster results.

As imaging systems and camera technologies improve, we see an increase in the resolution and size of images being collected, which leads to an increase in image file sizes and a challenge for model training. To use all of the information that is available, use the highest resolution images available. However, the limited amount of memory that is available on a single GPU can limit the size of a data set that you can use to train a model. During training, there is an expectation that the data, model definition, and calculated values all fit into the GPU memory.

To overcome this limitation, Watson ML CE includes Large Model Support. This function is available for Tensorflow, PyTorch, and IBM enhanced Caffe, and is used so that models can be built by using a combination of GPU and CPU memory space. Using this capability, it is possible to build more complex networks or use larger data points to train a DL model by swapping calculated tensors in and out of system memory until they are required for back propagation. The NVLink 2.0 connectivity between the GPU and CPU in the Power AC922 server ensures this can be done quickly without impacting heavily on training time.

The process for creating an accurate artificial vision model is highly iterative. Because models are trained from a data set, they should be tested for performance and bias before modifications are made to the data, the network design, or the hyperparameters to train the next iteration. With this iterative approach, you can build a model that is accurate enough to solve the business challenge without unwanted bias or error.

IBM PowerAI Vision makes it easy to deploy a model for testing within the GUI. The model creates an API interface that you can use to check the model against a test data set that is held back from the training data. You also can test the model within the interface by uploading test images to easily confirm that your model provides the expected results (Figure 3-1), and identifies any problems. IBM PowerAI Vision also provides a range of metrics for each trained model to help identify how data sets might need to be modified to provide better results.



Create Dataset

Start by adding images and video files to a data set.

Prepare Data

Label objects in images and video frames, then use auto labeling to finish adding labels to the whole data set; assign categories to images or videos, or label actions in videos.

Train Model

Select a few custom options to create your model.

Deploy Model

Deploy the trained model and receive an API link for an inference device.

Figure 3-1   IBM PowerAI Vision guides SMEes through the steps of model creation

It is also possible to deploy these trained models for long-term use so that they can be used in a production environment or connected to a specific imaging system. The models can be deployed by using IBM PowerAI Vision Inference Server, which provides the same API functions that are on the training system. These deployments are containerized, and provide a flexible and scalable method for continuous inference use.

Running the models on GPU-accelerated systems still offer the best performance, but systems without accelerated hardware can be used to reduce costs if performance is not the key requirement. Multiple models can be run on a single GPU or system, and updated versions deployed as needed without changing the applications connecting to them. This approach provides a resilient production-ready environment for inference separate from the training system, and has no impact on the training process for updating or creating enterprise artificial vision solutions.

The significant challenges for creating enterprise artificial vision solutions include:

- Finding the correct data to solve a problem.
- Effectively labeling the data.
- The time that is required to run multiple iterations of a model.
- Hardware limitations for image size or resolution.
- Testing and iterating a model.

The tools and capabilities that are available in various AI offerings on Power Systems serves can help overcome these challenges by supporting the data science teams to become more productive.

3.2 Natural language processing

Natural language processing (NLP) is still far away from the sci-fi technology behind the universal translators we see in movies. Language is full of nuances that even humans struggle to grasp.

As with any ML application, the main challenge is collecting the correct data in large enough amounts to train models, and then collecting the infrastructure resources to run the training job. Additionally, the nature of NLP solutions makes this challenge even bigger because there are still four open problems[1]:

1. Natural language understanding.
2. NLP for nonpopular languages. This problem is often called a *low-resource scenario*.
3. Large or multiple documents.
4. Data sets, problems, and evaluation.

These conditions call for data science professionals, including data scientists, data engineers, and data analysts, to pay close attention to the tools and infrastructure features that help overcome these challenges on the journey from proof of concept (PoC) to production.

Organizations generate enormous amounts of unstructured data in the form of text, and they have been doing so for many years. NLP researchers are working to come up with models that can adapt and generalize across different domains or new forms of data, where there is not necessarily much labeled data that is available. Most organizations want to implement NLP business solutions by using their own data sources. So, every time a new project requirement comes down the line, data science professionals must start from the beginning of the NLP AI workflow.

The data science team must create labeled data sets each time that a new NLP project starts on a new domain or even in a new language. With the current state of the art, data science teams cannot adapt existing models to new data sources. This situation, due to the iterative nature of the AI workflow, requires repeating the same tasks for each new language to create a data set to create data and algorithms:

- Find best practices from approachable research papers. Try to replicate them.
- Understand the data that is involved.
- Find reference code. Otherwise, a serious coding effort is required.
- Take the resulting models and apply them to their own data.

To overcome these challenges, the correct tools and infrastructure play an important role:

- Where is the needed data stored?
- What is the data subject?
- How do you quickly parallelize NLP data preparation tasks?
- How do you accelerate training for several projects concurrently?

Data science professionals should discuss with the LOB and IT colleagues how IBM Cognitive Systems can help address these challenges and answer these questions so that you can focus on the data science critical tasks behind data analysis, model building, and visualization.

One important requirement for data science teams working with NLP AI workflows is to interact with the data in a collaborative way. This function is important with NLP solutions where sharing new connections, pipelines, and models is fundamental to accelerating the NLP AI workflow and the NLP field of study.

[1] For more information about the NLP four open problems, see The 4 Biggest Open Problems in NLP.

IBM Spectrum Discover changes the way data science professionals interact with data by offering role-based access control (RBAC) so different teams can confidently share data across silos. Concurrently, and more critical to the NLP solutions team, is the speed with which text data sets can be created and labeled. IBM Spectrum Discover provides data insight for exabyte-scale unstructured storage. It provides a rich metadata layer on top of storage sources to enable your team to efficiently manage, classify, and gain insights from massive amounts of unstructured text data. You can use it to create custom tags and policy-based workflows to orchestrate deeper content inspection and activate text data in the AI workflow.

Accelerating the NLP AI workflow is key because of the lack of models that can generalize across multiple domains and the lack of labeled data sets. Your data science team needs an information architecture (IA) that can accelerate NLP data-preparation-specific tasks like removing stop words; tokenization; lemmatization; stemming and pruning; and model scoring, selection, and training. These goals can easily be attained by using IBM Spectrum Scale coupled to your organization's data lake and WMLA to provide the infrastructure to support an AI as a Service (AIaaS) model. Watson ML CE can also be considered when the project is just starting and there is no need for a shared infrastructure. Watson ML CE accelerates your AI development stack installation, including the most popular frameworks and libraries in a convenient distribution with optional IBM support (Figure 3-2).

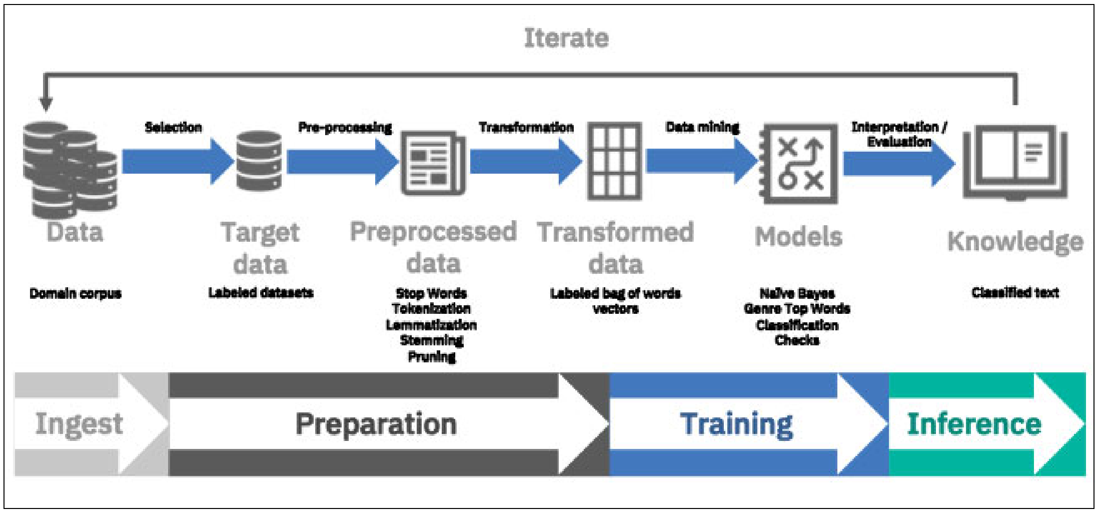


Figure 3-2 NLP AI workflow-specific tasks

Another side of data access is collaboration on data. With IBM Spectrum Scale, you can work on multiple projects across multiple sites with the same data. If collaboration is key, IBM Watson Studio becomes critical to build the NLP data science workbench where data scientists and data engineers can access data science tools, such as RStudio, Spark, Jupyter, and Zeppelin notebooks. Its intuitive user interface provides a collaborative project space for teams and individuals to reduce the time to value. Projects can contain notebooks, data assets, and collaborators.

One of the challenges data science teams face is to avoid becoming their organization's AI bottleneck as more projects require their skills and time. There are AI projects that do not need the expertise of a seasoned data professional. IBM Cognitive Systems offerings can help in those situations with tools to democratize access to AI. You can use these tools automatic ML and DL development so that you can run simpler NLP projects without heavy intervention from the data science team. They are also useful tools that the SMEs within the LOBs can use to help in the labeling stage of new domain or language text data sets.
H2O Driverless AI and IBM PowerAI Vision are the most useful tools in these situations:

H2O Driverless AI Although not as sophisticated as the manual approach, novice data science professionals and even LOB professionals can go through text data to find critical information to inform better business decisions. H2O Driverless AI sets the cornerstone for your AI information architecture to support your organization to go through the NLP AI workflow. It automatically converts text strings into features by using the most popular techniques like Term Frequency–Inverse Document Frequency (TFIDF), Convolutional Neural Networks (CNNs), and Gated Recurrent Unit (GRU). With built-in TensorFlow, H2O Driverless AI can also process larger text blocks and build models by using all available data to solve business problems like sentiment analysis, document classification, and content tagging. Version 1.7 includes the Bring Your own Recipes (BYOR) capability. With BYOR, data science professionals can quickly leverage critical NLP community knowledge by incorporating customizations and extensions to the platform. These customizations are Python code that is uploaded at run time, including:

- Custom ML models
- Custom scorers (classification or regression)
- Custom transformers
- Custom data sets

IBM PowerAI Vision When your NLP data science teams have limited skills, they can use other sources of data to create their text data sets, including print media like digital newspaper articles, magazines, blogs, and others. Images require them to apply DL techniques to scan images of text to be used for training your NLP model. IBM PowerAI Vision provides tools to simplify the process of creating image-based DL models.

3.3 Planning for the future

Data science professionals have a multi-faceted role in businesses today, often overseeing end-to-end workflows in AI-based projects from data ingest through model creation to model deployment and ongoing verification. With the increased pressure to use AI to provide the business with positive results, these professionals must make efficient use of their time. Common day-to-day predictive analytical tasks often iterate through the AI workflow:

- Prepare the data: Access the data from remote and local sources, and split the data between training and testing subsets.
- Train and deploy the models: Use the data feature engineering, model tuning, and model selection.
- Verify the results: Review for accuracy. Visualizations can be useful to help understand and provide interpretability (explanations) of the results.

Each of these tasks takes a measurable amount of time with serialization for an AI workflow requiring one step to complete before the next begins. Reducing any (or hopefully all) of these steps means that more time is spent on either larger numbers of workflows or development of higher-value capabilities.

Industries in which you can incorporate ML into analytics and decision-making are varied and can include insurance, energy and utilities, banking, government, manufacturing, retail, healthcare, and transportation.

IBM works with a leading automated ML vendor, H2O.ai, to incorporate their H2O Driverless AI product for automated machine learning (Auto-ML) with text and structured data. This enterprise-grade software can be used both by inexperienced and experienced data science professionals to accelerate their learning or expand their capabilities. It can also be used by subject domain users or LOB professionals to quickly provide results with no coding required.

- Data sets can be ingested from local files (many formats are supported) or by using remote data connectors for Hadoop Distributed File System (HDFS), Amazon S3, Google Cloud Store, and many other providers.
- Experiments are where H2O Driverless AI analyzes the data based on the selected target variable for results and automatically selects ML models based on included feature engineering recipes that are pre-built by expert data scientists. Expert-level customizations are selectable.
- There are three main knobs that can be used to adjust the outcome of an experiment. Adjusting one knob affects the others.
  - Accuracy
  - Time
  - Interpretability
- ML interpretability provides human-level understanding of the math and processes behind the ML model and its results.
- In addition to the GUI as the primary interface for using H2O Driverless AI, there are also deployable APIs for integration into automated processes and full explainability reports.

Some data scientists prefer to code, tune, and instrument their own ML models. IBM provides Watson ML CE for this purpose. Watson ML CE provides pre-compiled (both ppc64le and x86_64) versions of popular ML and DL frameworks. Watson ML CE was co-developed with the IBM GPU-enabled POWER9 processor-based server (Power AC922) to take advantage of the advanced hardware capabilities. For more information about the Power AC922, see *IBM Power System AC922 Technical Overview and Introduction*, REDP-5494 or *Cognitive Computing Featuring the IBM Power System AC922*, REDP-5555.

In addition, IBM Research™ developed an optimized, and enhanced drop-in library with accelerated ML libraries to speed processing on CPU and GPU in the Power AC922 that is called Snap ML. For more information, see 3.4, “Machine learning” on page 30.

Snap ML is delivered as part of the Watson ML CE Conda channel, which is freely accessible.

H2O Driverless AI is available from IBM and H2O. A trial can be acquired at Try Driverless AI.

Both Watson ML CE and H2O Driverless AI are available as installable software packages by using the native operating system methods for Red Hat Enterprise Linux Version 7.6 and Ubuntu Linux Server Version 18.04.1. Both are also available as Docker images: Watson ML CE from Docker Hub and Driverless AI from H2O.

For more information about, Watson ML CE, see the following resources:

- IBM Knowledge Center
- IBM Cognitive Systems developer portal

For more information about H2O Driverless AI, see the following resources:

- Using Driverless AI 1.7.1 documentation
- GitHub
- Try Driverless AI

3.4 Machine learning

As ML becomes more prevalent, there is increasing pressure on data science teams to build new solutions and capabilities by using data-driven ML techniques. Along with the increase in data volumes, the requests from businesses for innovation and new ML models is increasing. To meet growing demand, data science teams must overcome many challenges.

Data collection and preparation is a key first step in any data science project, including ML projects. You must identify the correct sources of data to help solve the business challenge. That data then must be organized and curated so that disjointed data can be linked together and easily manipulated by data science tools.

Using IBM Spectrum Discover, data can be searched from across multiple systems and services within a business by searching by file type, metadata entry, or custom tags. This feature makes it possible to pull out all of the files within a business that are labelled as a certain type or hold a particular type of data. With this feature, you can automatically add custom tags to files as they are created or processed, for example, tracking project usage for governance. This feature makes it easier to collate the data that is required for an ML project by easily finding the data that is most relevant to the business challenge.

There are benefits to be gained from collaborating on projects and sharing the resources, data, and expertise to jointly tackle a challenge. Watson Studio is designed as a collaborative workbench for data science teams to work together on multiple projects. You can share data sets and connections to external data sources, and collaborate on model building through notebooks, integrated development environments (IDEs), and visual modeling tools. Data science professionals can contribute to multiple projects with data sets and other assets that are managed within the projects to share or limit access to other teams as needed. There is notebook support for Python, R, and Scala, and RStudio support and tools for creating model flows by using visual interfaces.

Like other data science work, ML projects generally require a highly iterative approach. Data scientists work with a range of different ML algorithms, and work with the features that are available in the data to derive the greatest insight. This work might involve feature engineering approaches and identifying other sources of data to add to the working set. Each new combination of data and algorithm requires the training of a new model, which takes time to complete, so as you increase the number of options that you work with, the time requirements of the project also go up.

To minimize the impact of multiple iterations across the project, you can reduce the time that is required to train each model. Watson ML CE includes the Snap ML libraries. Snap ML adds GPU acceleration to some of the most common ML algorithms, speeding up training times up to 46x[2]. Snap ML introduces parallelization to the algorithms so that specialized accelerators like GPUs can improve performance, and improve scalability over devices or systems for larger data sets. These capabilities are available through standard API calls, and there is a SciKit Learn compatible interface for Python, which can be used to accelerate existing ML code with minimal changes. The algorithms that are accelerated by Snap ML include:

- Logistic regression
- Ridge regression
- Lasso regression
- Support vector machines
- Decision trees
- Random forests
- Gradient boosted machines

Other algorithms are being added, but if you use the Snap ML Python API, then any functions that are not accelerated by Snap ML fall back to the SciKit Learn versions for consistency. For larger training jobs, Snap ML can be distributed across GPUs and servers by using message passing interface (MPI) technologies or Spark runtime environments. Further performance improvements can be seen when using Snap ML on the Power AC922 server.

Distributing training across multiple devices and servers can also lead to shorter training times, allowing for greater numbers of iterations and the chance to build a higher-value model. By using WMLA, you can distribute your training runs across multiple systems and multiple GPUs in a fully managed and scalable way. This training can be run through Jupyter or Zeppelin notebooks or Spark jobs, and can elastically scale across the available resource to get faster results.

H2O Driverless AI from H2O.ai also helps find the most suitable algorithm for your data set. This tool automates many processes that are used by data scientists, and can train models based on multiple algorithms and engineered features to find the highest accuracy level. There are also tools to visualize data sets, help identify outliers or anomalies, or see where features are heavily correlated. These tools can help a data scientist to better understand the data set and also pick the most appropriate algorithm to use. H2O Driverless AI also uses GPU acceleration to speed up processing and speed results.

Another challenge in introducing ML algorithms into business processes is that of trust. Traditional methods of building or automating business processes are based around implementing rules, which creates processes that are easy to explain. With these processes, people can interpret how a decision was made by following the rules that implement that choice. With ML techniques, business processes can be built by using data-driven techniques that create effective models but cannot easily be explained.

H2O Driverless AI also includes interpretability functions, which you can use to take a complex ML model and explain how it makes decisions, including showing which features have the largest impact on the result. It uses ML techniques to show more easily understood linear models or decision trees in place of complex models that you can use to explain how decisions were made in individual cases, which can be shared with LOB professionals or used to confirm that no unwanted biases exist.

[2] For more information about Snap ML, see Snap ML.

The final challenge in an ML project is understanding how best to deploy a trained model for business use. This deployment should follow the iterative training process, including a period of testing to ensure the suitability of the model performance and that no unwanted bias or unfairness exist. At this stage, the model must be deployed to allow inference, usually as part of a larger application. As such, the ability to package up a trained model in a way that is easy to deploy, access, and use is valuable. Managing the model lifecycle is also valuable to track different versions of models as they are retrained on new data.

WMLA has model management and deployment functions. With IBM Watson ML Accelerator, you can quickly take a trained model and deploy it with a REST API service for developers to incorporate into applications. You can also use this function to quickly and easily test the models that you created. H2O Driverless AI also includes capabilities to export trained models for inference with run times for Python and Java that can be incorporated into applications.

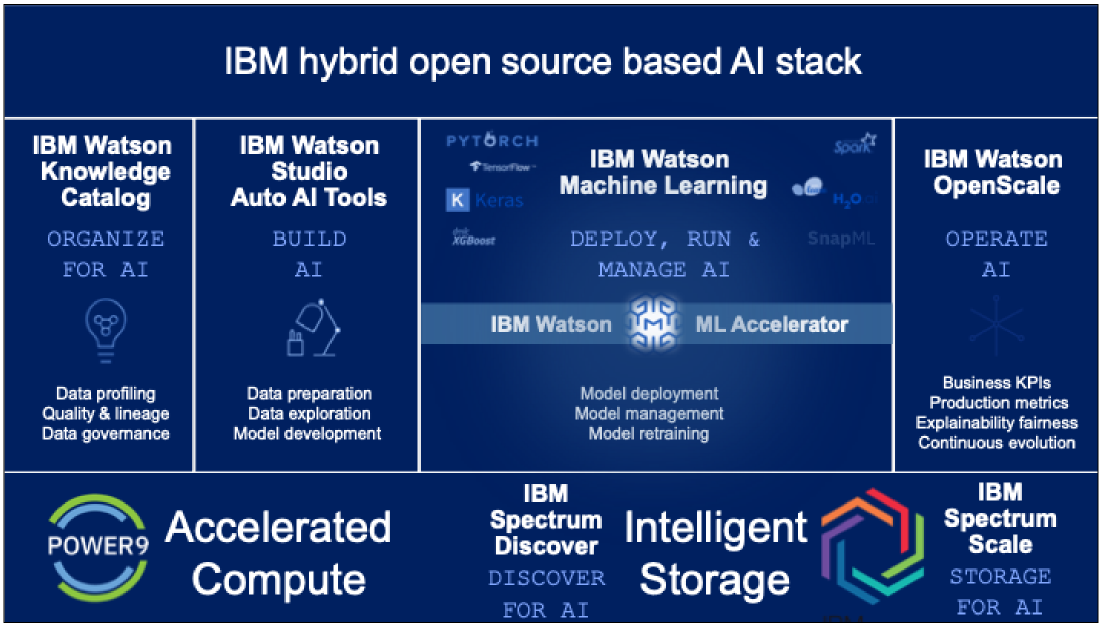
Throughout the development of an ML solution, there are techniques that you can use to get to higher levels of accuracy faster, from better understanding of the data through faster training iterations, and on to simplified model management, testing, and deployment.

3.5 AI teaming and collaboration

Data science is about teaming and collaboration to produce business results from enterprise AI solutions. The number of AI projects in organizations continue to proliferate, putting pressure on individual data science professionals. Additionally, related job roles that are focused on specific tasks in the AI workflow are appearing to distribute the load and provide focused efficiencies to speed time-to-value.

With the growth in the number of data science professionals, the ability to work together in a coordinated fashion and with efficient utilization of computing and storage resources becomes a key differentiator in a company's journey to AI. Rather than each data science professional getting their own physical servers or cloud instances, by using a unified workbench where tools, data, and models can be shared allows for reuse, sharing, and consistency.

IBM has an end-to-end AI strategy with tools for data management, model management, and ongoing production support (Figure 3-3).



IBM hybrid open source based AI stack

IBM Watson Knowledge Catalog
ORGANIZE FOR AI
Data profiling
Quality & lineage
Data governance

IBM Watson Studio Auto AI Tools
BUILD AI
Data preparation
Data exploration
Model development

IBM Watson Machine Learning
DEPLOY, RUN & MANAGE AI
IBM Watson ML Accelerator
Model deployment
Model management
Model retraining

IBM Watson OpenScale
OPERATE AI
Business KPIs
Production metrics
Explainability fairness
Continuous evolution

POWER9 Accelerated Compute

IBM Spectrum Discover
DISCOVER FOR AI

Intelligent Storage

IBM Spectrum Scale
STORAGE FOR AI

Figure 3-3 IBM hybrid open source-based AI stack

IBM Watson Studio provides a web-based developer tool bench that is built around the concept of projects. Collaborators can work in multiple projects according to their role: administrators, editors, or viewers. Within each project, there are assets with which to work:

- Data tools for preparation and visualization of the data (where appropriate). Data can come from local (on-premises) sources or remote (cloud) connections with built-in tools to prepare the data for ingest.
- Analytical assets like flows, notebooks, and ML/DL models. Models can be developed inside Watson Studio and sent for training and automated model deployment in production by using IBM Watson Machine Learning. Experiments on DL models can run in parallel to evaluate hyperparameters for the best results with your data and models.

Watson Machine Learning works with Watson Studio to provide the training, execution, and deployment of the actual ML and DL models. The primary interaction with Watson Studio and Watson Machine Learning is by using a graphical, web-based UI, but command-line and REST APIs are available for programmatic access.

WMLA is a Spark-based, multi-tenant offering that is designed to speed up ML and DL development on a cluster of GPU-enabled servers. The DL tasks can be self-contained or come from Watson Machine Learning or Watson Studio.

WMLA has many advanced features of interest to the data scientist looking to increase their efficiency:

- Distributed DL and GPU scheduling fully uses the limited GPU resources across the cluster, including:
  - Prioritization based on administratively defined policies between groups and users.
  - Preemption of GPU resources (if required) based on policy.
  - Dynamic addition or removal of GPU resources from running training jobs by using the Elastic Distributed Training (EDT) feature based on priorities or the fairshare resource allocation algorithm. With this capability, data science professionals do not need to know what the hardware foundation is. With just a few lines of code, they can use as many GPUs as needed without caring about where they are in the infrastructure.
- Real-time training visualization and runtime monitoring for unwanted conditions like divergence, overfitting, and underfitting. Upon notification, the user can stop the current training job when those conditions are detected and restart with the appropriate suggested hyperparameter values.
- Support for DL frameworks that come with Watson ML CE, such as TensorFlow, PyTorch, Caffe, Keras, Bazel, ONNX, and NVIDIA Rapids cuDF/cuML. You can bring your own framework and extend the capabilities of the tools.
- There are data ingest, preparation, and transformation tools that re based on Apache Spark, including splitting of data sets into training, validation, and testing subsets.

Real world example: Large US banks use Corporate Model Risk departments to monitor and maintain interpretability in ML and DL models to avoid fraud and comply with legal requirements. One such bank has thousands of models with massive amounts of data coming both from current transactions and historical data. They require the speed to run large subsets of data through these various models simultaneously while maintaining high levels of efficiency of GPU-enabled computational resources. By using the Power AC922 server with its fast CPU-GPU NVLink connection (moving large data through faster) and WMLA for efficient scheduling of the GPUs, they can achieve higher accuracies faster.

IBM Watson OpenScale focuses on the production operation of ML and DL models in an intelligent production environment:

- Ongoing verification of model accuracy and performance.
- Detects and mitigates model biases to highlight possible fairness issues.
- Explainability of transactions with a list of attributes that is used in decision-making and the weight of each attribute.

The Watson family of products that are mentioned in this section are available both in an on-premises Kubernetes-based cloud or off-premise cloud. Multiple vendors clouds are supported. A no additional charge trial and more information can be found at the following locations:

- IBM Watson Studio
- IBM Watson Machine Learning
- IBM Cognitive Systems developer portal
- IBM Watson OpenScale

4

Point of view: Information technology

The use of artificial intelligence (AI) has spread throughout organizations in both large and small projects. Historical implementations of computing power were under the management of information technology (IT) departments. Today, in a race to quickly implement enterprise AI solutions that are used by data science professionals, the various lines of business (LOBs) bypass the traditional IT organization with their own purchase, implementation, and support of systems (or cloud instances). Rather than waiting on what is perceived as sluggish response times from IT, the LOBs have created their own miniature IT organizations, which are sometimes called *shadow IT*.

This shift created an increased level of complexity for companies who end up with disjointed islands of compute resources, each possibly with their own security policies, software dependencies, enterprise support contracts, and so on. Situations that can cause challenges for IT include:

- LOB-purchased resources that are already used in production are transferred to the IT organization for on-going maintenance and support. Often, those resources do not adhere to the guidelines and policies that are defined by IT, so exceptions must be requested and more skills acquired.
- Cloud-based resources that the IT organization must support add another conflict with policies, procedures, and security, and the hardening of those cloud assets might not be as easy as the on-premises resources that are managed by IT.

For many new workloads, IT organizations have lost their position of prominence when it comes to selecting the tools and infrastructure for strategic projects. They are mostly relegated to managing the existing infrastructure. This trend is being challenged by IT leaders who align closer to the needs of the business while those leaders who are not making that challenge continue to lose the budgets that are associated with the support of the infrastructure that is needed for these projects. That money now goes to shadow IT projects, which do not benefit from the oversight or experience of the IT department. To bridge this gap, IT departments must adapt their services and offerings to meet the new demands of the business.

© Copyright IBM Corp. 2019. All rights reserved.

The cycle of computing swings like a pendulum between strict management and consolidation, and the untethered distribution of compute resources. Companies typically have compute resources in both places: LOBs and traditional IT. Supporting the business goals of reducing costs, simplifying management, and reducing complexity, many organizations are moving to standardize their IT environments and deliver consistency across the LOBs.

For some companies, this might mean pushing everything into the cloud with a prescribed list of cloud vendors and a consolidated cost structure. For other companies, it might mean a consolidation of data centers with greater use of virtualization and private clouds to increase the usage and flexibility of compute resources. As before, the end goal might be a mix of both strategies with an eye on intelligent placement of data and workloads where it makes sense.

The IT organizations of today are typically concerned with the following considerations when talking about the implementation of AI initiatives with the LOBs and data science teams:

- Where is the data for AI model training regarding the GPU-enabled compute resources that are required?
  - How much data? How far does it have to move (transfer)?
  - Are multiple copies needed, and if so, how many and what are the security and retention policies?
- Where are the best locations for development versus production workloads? Using hybrid cloud is a key method for implementation to get the best of both worlds (on- and off-premises). This item often involves determining the best infrastructure for your AI workloads within the available budget.
- For training, how can you best use the limited GPU-enabled compute resources for maximum efficiency?
- After training, where is the best location for inferencing? Where does that data come from, and what are the security and retention policies for that data?

These items *support* the business objectives, but *not instead* of the business objectives.

In summary, the IT organization is there to support the LOBs and the data science teams to produce the best in business outcomes from their enterprise AI solutions. They are a partner in the success of the company and not an antagonist to “work around”. There is no better organization to deal with enterprise IT requirements than the IT department.

In this chapter, we describe five use cases to understand how IT can address the business requirements for implementation of AI initiatives by creating solutions that are based on IBM Cognitive Systems AI offerings.

4.1 Artificial vision

Advances in artificial vision technologies are creating challenges for IT departments because different LOBs want to implement these capabilities. Data science teams require access to the latest tools, frameworks, and libraries, and a more flexible service so that they can test multiple approaches with different parameters. It is up to the IT teams to provide these technologies rapidly and also focus on key factors such as:

- Security of the data that is used and the systems.
- Resiliency of the service and the data.
- Maintenance and upgrades to keep current.
- Integration with other systems and data stores.
- User management and access controls.

To balance these factors with the requirements of the data science teams for new tools and functions, IT departments might require new deployment methods or management techniques. A common approach is to introduce cloud-like capabilities so that data scientists can quickly create environments to label images, train models, and test their output. The burden on the IT administrators then moves from ongoing maintenance of running environments to defining services and images that can be rapidly deployed in a predictable and automated way.

To build artificial vision solutions, there are many steps that data science teams must iterate through to build a high accuracy model:

- Data collection and curation
- Labeling of data
- Model training
- Testing of a trained model
- Deployment for production

Each step has different requirements for the underlying technology and the way it is configured, and in some cases multiple systems or environments are used concurrently. The IBM PowerAI Vision toolset includes tools to accelerate and simplify each step of this process, providing for rapid development of various deep learning (DL) models for artificial vision. The whole software stack is fully containerized and can be deployed in different ways, depending on how to best integrate with other IT systems.

The following sections describe each of the possible solution.

Stand-alone system installation

IBM PowerAI Vision must run on GPU-accelerated systems like the IBM Power Systems AC922 server. It is possible to install the software on a single system and use all the resources that are available to run every aspect of the model creation process. Model training or deployment can use GPU resources as needed if they are not already in use. The installation process includes the creation and configuration of a Kubernetes-based container orchestrator; no Kubernetes skills are required. This Kubernetes instance manages all resource allocations and mapping of container services.

Private cloud installation

It is also possible to run IBM PowerAI Vision on a previously installed Kubernetes cluster that has GPU resources within it. There is a publicly available Helm chart that specifies the containers and dependencies that are required to deploy a working instance of the software stack. This stack can then be deployed through private cloud management systems like IBM Cloud Private or Red Hat OpenShift, where multiple instances of IBM PowerAI Vision can run on a single cluster while using all available resources, including GPUs. With this method, other workloads can use GPU resources when they are available.

With either method, it is possible to integrate with existing storage systems for long-term persistent storage of data sets and trained models. It is also possible to import image data from various different sources, either directly through the simple web-based interface or through the REST API that is built into IBM PowerAI Vision. Security policies can also be applied at the server level, cluster level, or network level to restrict access. User access can be defined and managed to limit access to the system and data.

Using a containerized approach makes it easier to deploy upgrades and perform maintenance tasks during the life of the software deployments. To gain new features and capabilities rapidly, you can add updated container images to the systems restart the service. In a private cloud environment, all running containers can be ported to other systems if needed for server maintenance.

The training aspect of DL development is highly iterative, and uses specialized hardware accelerators like GPUs to improve performance. Given the nature of this type of workload, there is little need for a high availability strategy for the training runs because they can be restarted if needed if there is a failure. Data resiliency of the labeled data sets is more important, and so within IBM PowerAI Vision it is possible to export labeled data sets to external systems, either from the web interface or through the APIs.

The resiliency requirements are typically different when moving a trained model into production deployment for inferencing because of the nature of the DL model and how the importance to the business can require a more resilient approach. With the IBM PowerAI Vision Inference Server, you can deploy trained models in a separate environment from the training environment. This deployment is also container-based, and you create and run container instances with inference services exposed as APIs that are ready for applications to use them. These container deployments are scalable, so you can run multiple independent instances of the same model with a load balancer presenting them as a single endpoint to applications for either performance or redundancy reasons.

It is also possible to deploy multiple models onto a single GPU or a single system to get more efficient use of the hardware resources that are available. The IBM PowerAI Vision Inference Server deployments can use GPU accelerators when they are available, but can also run as CPU-only containers, although at a reduced performance.

For data science teams that use other tools to build artificial vision capabilities, it is also possible to run the common DL frameworks and tools in both baremetal- and container-based installations. IBM Watson Machine Learning Community Edition can be installed on a single system, or a cluster of systems with GPU accelerators.

The following sections describe the possible installations.

Single-system installation

By using this installation, you have access to the common DL frameworks like Tensorflow, PyTorch, Caffe, and Keras on a single system, through the command line, Jupyter Notebooks, or integrated development environments (IDEs). Jobs can be run on all of the resources within the single system.

Clustered installation

You can use multiple systems to distribute training workloads across many servers. The Distributed Deep Learning (DDL) function spreads the model of a training job across any number of GPU accelerators within the clustered environment by using a shared file system for simultaneous data accessibility, which enables faster training times because parallel tasks are shared across the resources that are available.

Private cloud installation

All of the frameworks and tools can be run within containers, with multiple containers being used for DDL if needed. Users can request specific numbers of GPUs for their workloads and run them in a more cloud-like environment. This setup enables a level of isolation while also adding flexibility to the deployments.

For collaborative environments, IBM Watson Machine Learning Accelerator (WMLA) simplifies some of the common tasks of creating data sets and training models while still providing a high level of control over the process to the data scientists. This setup can be deployed on any size cluster of systems with GPU accelerators, and uses Elastic Distributed Training (EDT) to maximize the use of the resources that are available across all submitted training jobs. This software also manages user and role-based access control (RBAC) in multi-user environments so that limits can be placed on users or projects, and controls placed on data access.

IBM PowerAI Vision and WMLA both include support from IBM. When these tools are run on the Power AC922 server with NVIDIA Tesla V100 GPU accelerators and use either Ubuntu Linux or Red Hat Enterprise Linux, it is possible to get a hardware and software solution that is fully supported by IBM, including the underlying open source frameworks and libraries. This solution provides greater confidence in the offerings, and helps with your deployment should you need it. There is also an optional support offering for the Watson ML CE software bundle that is available if required, although the standard offering is a no-charge product with community support only.

The AI offerings on Power Systems servers aim to simplify the deployment and implementation of appropriate software across the hardware resources available while allowing IT teams the control they need to implement their own security, resiliency, and management policies.

4.2 Natural language processing

Natural language processing (NLP) is the set of technologies and methodologies from linguistics, computer science, information engineering, and AI that deals with how computers process and analyze natural language data.

Business solutions that are based on these technologies and methodologies include document classifiers, cognitive optical character recognition (OCR), voice-of-customer, advertisement optimization, and virtual assistants. For a brief description of each business solution, see 2.2, “Natural language processing” on page 15.

This publication tries to bridge the gap between LOBs, data science, and IT so they can work together to push AI business solutions from proof of concept (PoC) to production at scale. There is no better organization to deal with enterprise IT requirements than the IT department.

IT departments can support NLP solutions at scale by making the organizations' information readily available to start creating training data sets. To support the more iterative nature of an NLP AI workflow, as described in 3.2, "Natural language processing" on page 26, IT must create an information architecture (IA) to cycle through the four stages of AI development in the shortest possible time and at the optimal cost to gain the most reliable AI models.

The NLP data science team needs an IA ready for collecting and generating text data for projects. Unstructured text data can come from multiple sources, so IBM Spectrum Discover helps your NLP solutions development team rapidly find the data that they need by understanding what is contained in your data lake, where it comes from, and how accurate it is. In addition, the IT teams must support enterprise requirements like:

- RBAC so that departments can securely share data across silos, which enables NLP data science teams to discover and access the data that they need.
- Governance to comply with access restrictions so the correct people have access to the correct data at the correct time without fearing a breach of compliance security.

IBM Spectrum Discover ensures that every piece of information is securely indexed, classified, accessible, and governed.

Data science professionals dealing with NLP solutions must parallel process stemming, counting, creating frequency vectors, and other tasks repeatedly on each domain or language data set. IBM Spectrum Scale coupled with the parallel processing of Hadoop can maximize I/O in the preparation stage, maximize throughput and minimize latency in the training stage, and minimize latency in the inference stage. Concurrently it can minimize the movement that is required between stages to speed up end-to-end cycles. To be aligned with business priorities, it can maintain a cost-optimized archive to reduce overall storage costs.

IT departments must collaborate with the data science teams to decide which solution provides the most suitable capabilities based on the data science skills that are available. They can select an auto AI platform to support the NLP AI workloads with the smallest possible footprint by implementing H2O Driverless AI and IBM PowerAI Vision. You can use these tools for auto machine learning (ML) and auto DL for simpler NLP projects without heavy intervention from the data science team. H2O Driverless AI and IBM PowerAI Vision are excellent tools to allow subject matter experts (SMEs) at the LOB level to go through the AI workflow without help and to collaborate with the experienced data science team in the labeling stages of new domain or language data sets.

Finally, as more data science projects are started across a business, the IT team might need to implement a reference architecture so that their organization can offer an ML as a service environment. As data volumes grow into the petabyte range and there are multiple teams and multiple AI projects running concurrently, tools like WMLA are recommended. It provides supported and optimized versions of the leading open source AI frameworks in an integrated and supported package. It also includes enhancements to help address large and complex NLP projects. Customers might start small with a single node for a single experienced data scientist; in that case, consider using Watson ML CE. IBM Watson ML CE accelerates your AI development stack installation including the most popular frameworks and libraries in a convenient distribution with optional IBM support.

4.3 Planning for the future

One common area where businesses are using ML and DL techniques is in the traditional area of predictive analytics. The ability to answer the question of “what is the next best course of action based on these past data points?” can be critical to the success of a business process, product, or economic outcome. ML can add accuracy to these traditionally data- and compute-intensive tasks. The typical AI workflow involves:

- Data ingest and formatting: Selecting the correct data to analyze from local, shared, and remote sources. Modifying the data for correct usage by the tools might require multiple copies if inline modifications are not possible. Typically, this requires much data, multiple copies, and intelligent selection of data.
- Building the models: A typically iterative process requiring specialized expert selection of models, feature engineering, and modification of the tuning parameters to get to the wanted results. This process often streams the data through the models by using CPU and GPU resources for the ML processing.
- Verification of the results for accuracy: Visualizations are helpful, and an explanation of how the results were derived is often needed to remove the “black box” opinion of AI models.

Providing an IT infrastructure that can meet the needs of the AI workflow is the goal of the IT organization. It includes the software tools and hardware that are required to meet the business goals expeditiously.

Every organization runs on data. Providing correct and relevant data as input for an ML analytical pipeline is key to getting accurate results. IBM Spectrum Scale is a high-performance, shared file system with advanced capabilities that include integration of traditional POSIX-file, object, CIFS, and NFS into one global namespace. IBM Spectrum Scale is used on the most demanding supercomputers on the planet with hundreds of petabytes of storage, billions of files, and hundreds of gigabytes per second of throughput. It is designed to meet the needs of the largest data stores, and that same technology can be implemented at a customer’s location. Most importantly, it can be integrated with a customer’s existing storage if needed, or use a building block approach of software-defined storage (SDS) components with advanced reliability and availability capabilities to grow a new data hub.

Using the correct data often means knowing information about your data. IBM Spectrum Discover connects to file and object storage (both on-premises or in the cloud) to ingest, consolidate, and index those data sources. Data science professionals can catalog, query, and gain insights from large amounts of unstructured data by selecting the correct data for predictive analytics.

IBM and H2O.ai brought their flagship automated ML product H2O Driverless AI to the Power platform. Junior data scientists and SMEs can use H2O Driverless AI with no coding that is required to achieve high predictive accuracy similar to results that are achieved by advanced data scientists. H2O Driverless AI uses automated model selection and feature engineering based on the data that is available. Some of the algorithms include linear models, neural nets, clustering and dimensionality reduction models, and many traditional approaches like one-hot encoding. More experienced data scientists and analysts can adjust details to fine-tune the results.

H2O Driverless AI uses both GPUs and CPUs depending on the model and available hardware resources. When used on Power AC922 servers, H2O Driverless AI can run on baremetal with NVIDIA CUDA 10.0 or later drivers that are installed on Red Hat Enterprise Linux Version 7 or Ubuntu Version 16.04. For customers who are focusing on containerization, a Docker container is also available that supports GPUs if you use the nvidia-docker2 version of the NVIDIA Container Toolkit. Customers already using WMLA for efficient usage of their Power AC922 cluster can integrate and start H2O Driverless AI instances by using the WMLA scheduler.

For customers who develop their own ML models with popular frameworks like TensorFlow, PyTorch, NVIDIA cuML, and more, Watson ML CE is provided. Watson ML CE is available through a public Conda channel for use on a per-user basis running baremetal Red Hat Enterprise Linux Version 7.6 or Ubuntu Version 18.04.1 with Anaconda Version 2019.03. Watson ML CE is also available in various Docker images with a selection of built-in frameworks (single frameworks or all frameworks) and a CPU or GPU selection for both Python2 and Python3 from Docker Hub. For more information, see Docker Hub.

Included with Watson ML CE is an accelerated ML library that is available only for ppc64le architectures: Snap ML. Snap ML supports popular ML models like logistic regression, linear regression, support vector machine, and decision tree or random forest classifiers.

Some of the frameworks can take advantage of an advanced distributed mode that is called DDL, which provides near-linear scaling of training times across a cluster of up to four IBM GPU-enabled servers. DDL takes advantage of low-level remote memory direct access (RDMA) communications that are possible with InfiniBand networks that use traditional HPC protocols that are provided by IBM Spectrum MPI (included). Larger clusters (more than four nodes) are possible with WMLA in addition to Watson ML CE. Customers can use DDL with traditional Ethernet technologies, but the lower overall bandwidth and higher impact with the full TCP/IP stack does not provide as great a performance boost.

For more information about the Power AC922 server, see *IBM Power System AC922 Technical Overview and Introduction*, REDP-5494 or *Cognitive Computing Featuring the IBM Power System AC922*, REDP-5555.

H2O Driverless AI is available from IBM and H2O.ai. To obtain a trial version, see Try Driverless AI.

For more information and documentation about Watson ML CE, see the following resources:

- IBM Knowledge Center
- IBM Cognitive Systems development portal

For more information and documentation about H2O Driverless AI, see the following resources:

- Overview - Using Driverless AI 1.7.1 documentation
- Driverless AI at GitHub

4.4 Machine learning

As ML becomes an increasingly common workload across businesses, data science teams are more important than ever. With a focus on deriving data-driven solutions, their requirements for IT infrastructure and administration provide different challenges than the ones for traditional IT. Data is the core of ML, so the management of data access is important, but so is user access as you manage multiple projects across your infrastructure. For a multi-user ML environment, you must consider:

- User management and integration to authentication.
- Data management and access control lists.
- How the users access the resources.
- What tools and applications are available.
- How to ensure the most efficient use of resources.
- Collaboration requirements of the data science team.

For ML workloads, the requirements of the data science team might be diverse. There are a range of common frameworks and libraries that might be required, and with a rapid pace of development there might be requests for different versions. Most of the work is carried out in a Python environment, and access may be requested through the command line, web-based notebooks, or IDEs. There might be dozens of different variations of the environments that are requested.

Watson ML CE makes the process of creating and managing environments easier by using the Anaconda Python package and environment manager. With it, users can create their own Python environments and choose which version of Python to use, and which extra packages are installed. From the public IBM repository, they can download and install optimized versions of common frameworks and packages, specifying the versions as required. The tools pull in all of the needed dependencies, and allow multiple environments to exist independent of each other so that multiple versions can be used concurrently.

Watson ML CE also includes Snap ML, which is a set of GPU-accelerated libraries for common ML algorithms. When running on GPU-accelerated servers like the Power AC922 server, it can provide accelerated training times by running threads in parallel to use GPU capabilities. Because Snap ML includes an API service that is compatible with the commonly used SciKit Learn library, data scientists can use this acceleration with minimal changes to their code, which increases the speed of training and the overall performance of creating an ML model.

These capabilities are also available in WMLA, a multi-user environment for data science teams. It can be run across multiple systems, and it provides the capabilities to manage users, data, and access across the whole cluster. With integration into existing user management systems like LDAP and respecting access control lists from shared file systems, WMLA can integrate into your existing systems and offer a highly flexible data science platform.

WMLA can also increase the efficiency of resources because it acts as a scheduler for DL and ML workloads. Users can access the environment through various methods, including Jupyter or Zeppelin Notebooks, command line access, or a web-based interface. However, all jobs are run through the scheduler, which can allocate resources in a flexible manner to increase throughput. This task is done elastically so that new jobs run interactively while others continue training. Users can choose which environments and packages with which they work, which offers flexibility in how they work, with GPU acceleration that is available for the workloads that benefit from it.

WMLA also has a strong focus on systems management and administration by collecting and presenting metrics for system performance, usage, and health. It can record user or group usage of the resources so that you can cap some teams or give priority access for others. Policies can be defined to ensure that the available resources are being used in the most effective way, and usage metrics can be used to create chargeback or showback reports for different teams or departments.

Further collaboration can be achieved by using Watson Studio on the systems. It provides a data science workbench so that project teams can share access to data sets, notebooks, and other tools to collaborate on their work. Any training jobs that are created from this work can be deployed onto a WMLA cluster to make the most efficient use of resources and ensure that everything is managed and monitored effectively. With the collaboration tools, data science teams can share their expertise, data, and code, which can lead to improved results. Watson Studio also includes tools to simplify many of the steps of the AI workflow, from data preparation through graphical model definitions to rapid deployment of training jobs.

Another tool that simplifies the process of creating ML models is H2O Driverless AI. It provides automation of many stages of the AI workflow, which reduces the burden on a data scientist, and enables testing of a wider range of algorithms to get the highest accuracy level. These tools can be deployed within an existing cluster of systems running WMLA, sharing resources as needed with other workloads. It is also possible to install it on single systems as an independent workload.

Because H2O Driverless AI uses GPU acceleration systems like the Power AC922 server with NVIDIA Tesla V100 GPUs, it is recommended to get the best performance. The software can be installed on the base Linux operating system, or as a Docker container so that it can be deployed in a private cloud environment. Multiple containers can be run simultaneously with GPUs allocated to each container in whole number increments.

ML workloads introduce new requirements for the IT teams within businesses. These requirements are addressed by the various AI offerings on Power Systems servers to ensure that the data science teams can work effectively while applying security, access, and management policies.

4.5 AI teaming and collaboration

Growth in the area of AI projects within organizations often occurred in an unplanned fashion with separate projects cropping up throughout the organization. Data science professionals working on these projects have not necessarily communicated with each other, and use their own individual or departmental computing and storage resources. IT departments are challenged to provide a superior service to those data science teams and LOB that meet the following requirements:

- Securely provide access to the data that is required to train ML and DL models at a speed and volume to meet the business needs.
- With prioritization based on business needs, provide the necessary compute (both CPU and GPU) resources for use by different teams or individuals.
- Make available machine and DL tools, libraries, and associated programs that are familiar to the data scientists, and meet or exceed performance requirements.

- Provide cloud-like efficiency at a reasonable cost:
  - Rapid availability
  - Simple and centralized management
  - Flexibility to respond to surges in demand
- Provide an IA to broaden the access of LOB professionals to AI tools that can involve the professionals in the AI workflow (IBM PowerAI Vision and H2O Driverless AI).

New applications are often built to use cloud-like object storage, but historical data in enterprises typically is in traditional file-based repositories. IBM Spectrum Scale is an SDS solution that can use both existing storage and unified file-and-object storage integration. IBM Spectrum Scale provides the storage for the most demanding implementations on the planet (at the time of writing): the Summit supercomputer at Oak Ridge National Laboratory with 250 PB of storage that can sustain 2.2 TBps of sequential throughput. Encryption, parallel backups, quality of service, and full auditing support are a few of the features that are available in IBM Spectrum Scale that can support a large amount of data and an extensive number of data science projects.

IBM Watson Studio and IBM Watson Machine Learning provide a web-based data science tool bench that provides data preparation tools, data visualization tools, model building, management, and more. From these products, GPU-accelerated training jobs can be sent to WMLA running on a cluster of Power AC922 servers. The IT organization can provide a private cloud-based infrastructure running Kubernetes for IBM Watson Studio and IBM Watson Machine Learning, and a baremetal cluster for WMLA.

WMLA is built on IBM Spectrum Conductor®, a Spark-based and multi-tenant cluster offering that provides tools for data ingestion and manipulation, and Deep Learning Impact, which supports DL training and inference with advanced features. Installing WMLA on a cluster of Power AC922 servers running Red Hat Enterprise Linux Version 7.6 requires a shared file system, and IBM Spectrum Scale provides that capability. There is no need to create the Apache Spark cluster from scratch because that task is handled by the WMLA installer.

WMLA has two features that are designed to use multiple GPUs (on the same node or across more than one node) to reduce the overall training time and reduce latency in the inferencing stage: DDL and EDT.

- DDL uses a popular high-performance computing message passing protocol that is called the message passing interface (MPI). IBM Spectrum MPI (included with WMLA) provides reliable messaging, typically by using low-latency, high-bandwidth networks like InfiniBand. If InfiniBand is not available, traditional Ethernet can be used but at reduced performance compared to similarly rated InfiniBand. Ethernet speeds of 40 Gbps or 100 Gbps are recommended. With WMLA, all the Power AC922 nodes in the cluster can participate, which provides near-linear scaling in training times.
- EDT also uses the network for communication, but relies heavily on the shared file system for subtask synchronization between GPU processes. Using a high-speed shared file system like IBM Spectrum Scale can be beneficial. IBM Spectrum Scale supports tiered storage such as tier 0, which is composed of Flash or NVMe memory with tier 1 based on SSD technology and tier 2 based on hard disk drives. EDT can dynamically add or remove GPU subtasks from the training job based on the queue workload and administratively defined priorities between users and groups.

WMLA provides access to many of the latest ML and DL frameworks for the data science professionals to use from the embedded Watson ML CE. These frameworks include current releases of TensorFlow, PyTorch, Caffe, Keras, Bazel, ONNX, NVIDIA Rapids cuDF/cuML, and more. Watson ML CE also includes accelerated ML libraries for the Power AC922 server, which collectively are called Snap ML.

Most importantly for IT and the data science teams, these frameworks and the advanced DL capabilities of WMLA (like the Spark configuration, EDT, and hyperparameter optimization) are all fully supported by IBM. If a data scientist wants, for example, the Spark framework of WMLA can be used to add more frameworks that are not included with Watson ML CE.

For more information the topics in this section, see the following resources:

- IBM Watson Studio
- IBM Watson Machine Learning
- IBM Cognitive Systems development portal

5

Conclusion

This publication helped the line of business (LOB), data science, and information technology (IT) professionals find common ground and work as a team to build an enterprise artificial intelligence (AI) solution.

These professionals provide the following items:

- The business challenge that is backed by a business case.
- The AI techniques that are needed to answer those business questions.
- The information architecture (IA) to support the AI workflow from proof of concept (PoC) to production at scale.

Table 5-1 shows a list of IBM Cognitive Systems offerings and capabilities at the intersection of a persona and a popular use case.

Table 5-1 IBM Cognitive Systems capabilities

Persona / Use case	Line of business	Data science	Information technology
Artificial vision	IBM PowerAI Vision	IBM Spectrum Discover IBM PowerAI Vision Large Model Support	IBM PowerAI Vision IBM Watson Machine Learning Community Edition IBM Watson Machine Learning Accelerator (WMLA)
Natural language processing (NLP)	IBM Spectrum Discover	IBM Spectrum Scale IBM Spectrum Discovery Snap ML WMLA H2O Driverless AI	IBM Spectrum Scale IBM Spectrum Discovery WMLA
Planning for the future	H2O Driverless AI	H2O Driverless AI IBM Watson ML CE Snap ML	H2O Driverless AI IBM Watson ML CE Snap ML

© Copyright IBM Corp. 2019. All rights reserved.

Persona / Use case	Line of business	Data science	Information technology
Machine learning (ML)	Watson Studio Watson ML CE H2O Driverless AI	IBM Spectrum Discover IBM Watson ML CE WMLA Snap ML	IBM Watson ML CE WMLA Snap ML H2O Driverless AI
AI teaming and collaboration	IBM Watson Studio IBM PowerAI Vision WMLA IBM Spectrum Scale	IBM Watson Studio WMLA IBM Watson OpenScale	IBM Spectrum Scale IBM Watson Studio WMLA

AI is different in many ways than the IT solutions that were previously implemented in organizations. It requires a new vocabulary, new skills, and new ways of working to succeed. Different parts of the business must work together to tackle the challenges of bringing enterprise AI solutions up to the level of your other mission-critical solutions.

After they finish this publication, LOB, data science, and IT can sit around the table and have meaningful conversations by using a common language and set of tools to discover together how enterprise AI can be put to work to propel their business with competitive advantage and create sources of revenue.

Related publications

The publications that are listed in this section are considered suitable for a more detailed description of the topics that are covered in this paper.

IBM Redbooks

The following IBM Redbooks publications provide more information about the topic in this document. Some publications that are referenced in this list might be available in softcopy only.

- *Cognitive Computing Featuring the IBM Power System AC922*, REDP-5555
- *IBM FlashSystem A9000 and A9000R Architecture and Implementation (Version 12.3.2)*, SG24-8345
- *IBM PowerAI: Deep Learning Unleashed on IBM Power Systems Servers*, SG24-8409
- *IBM Power System AC922 Technical Overview and Introduction*, REDP-5494
- *IBM Power Systems LC921 and LC922: Technical Overview and Introduction*, REDP-5495

You can search for, view, download, or order these documents and other Redbooks, Redpapers, web docs, drafts, and additional materials at the following website:

ibm.com/redbooks

Online resources

These websites are also relevant as further information sources:

- H2O Driverless AI

  https://www.h2o.ai/products/h2o-driverless-ai

  http://docs.h2o.ai/driverless-ai/latest-stable/docs/userguide/index.html

  https://github.com/h2oai/tutorials/tree/master/DriverlessAI
- IBM Cognitive Systems developer portal

  https://ibm.biz/poweraideveloper
- IBM Spectrum Discover

  https://www.ibm.com/us-en/marketplace/spectrum-discover
- IBM Spectrum Scale

  https://www.ibm.com/us-en/marketplace/scale-out-file-and-object-storage
- IBM Video Analytics

  https://www.ibm.com/support/knowledgecenter/en/SSKRA3_1.0.0/va/kc_welcome.html
- IBM Watson Knowledge Catalog

  https://www.ibm.com/cloud/watson-knowledge-catalog

© Copyright IBM Corp. 2019. All rights reserved.

- IBM Watson Machine Learning Accelerator

  https://ibm.biz/poweraidevelopment
- IBM Watson Machine Learning Community Edition CE

  https://www.ibm.com/support/knowledgecenter/en/SS5SF7_1.6.1/welcome

  https://ibm.biz/poweraidevelopment
- IBM Watson OpenScale

  https://www.ibm.com/cloud/watson-openscale
- IBM Watson Studio

  https://www.ibm.com/cloud/watson-studio
- Try H2O Driverless AI

  https://www.h2o.ai/try-driverless-ai/

Help from IBM

IBM Support and downloads

ibm.com/support

IBM Global Services

ibm.com/services

REDP-5556-00

ISBN 0738458058

Printed in U.S.A.

Get connected

Redbooks®

ibm.com/redbooks