IBM Power Systems
Enterprise AI Solutions

Glen Corneau
Andrew Laidlaw
Marcos Quezada
Note: Before using this information and the product it supports, read the information in “Notices” on page v.

First Edition (August 2019)


This document was created or updated on August 23, 2019.

© 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

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Notices</td>
<td>v</td>
</tr>
<tr>
<td>Trademarks</td>
<td>vi</td>
</tr>
<tr>
<td>Preface</td>
<td>vii</td>
</tr>
<tr>
<td>Authors</td>
<td>vii</td>
</tr>
<tr>
<td>Now you can become a published author, too!</td>
<td>viii</td>
</tr>
<tr>
<td>Comments welcome</td>
<td>viii</td>
</tr>
<tr>
<td>Stay connected to IBM Redbooks</td>
<td>ix</td>
</tr>
<tr>
<td><strong>Chapter 1. Introduction</strong></td>
<td>1</td>
</tr>
<tr>
<td>1.1 Enterprise AI</td>
<td>2</td>
</tr>
<tr>
<td>1.1.1 AI workflow</td>
<td>2</td>
</tr>
<tr>
<td>1.2 IBM Cognitive Systems offerings</td>
<td>4</td>
</tr>
<tr>
<td>1.2.1 AI cognitive infrastructure</td>
<td>4</td>
</tr>
<tr>
<td>1.2.2 Organize for AI</td>
<td>5</td>
</tr>
<tr>
<td>1.2.3 <strong>Build AI</strong></td>
<td>6</td>
</tr>
<tr>
<td>1.2.4 Deploy, run, and manage AI</td>
<td>8</td>
</tr>
<tr>
<td>1.2.5 Operate AI</td>
<td>9</td>
</tr>
<tr>
<td>1.3 How this Redpaper is organized</td>
<td>10</td>
</tr>
<tr>
<td>1.3.1 Selected AI use cases</td>
<td>10</td>
</tr>
<tr>
<td><strong>Chapter 2. Point of view: Line of business</strong></td>
<td>13</td>
</tr>
<tr>
<td>2.1 Artificial vision</td>
<td>15</td>
</tr>
<tr>
<td>2.2 Natural language processing</td>
<td>16</td>
</tr>
<tr>
<td>2.3 Planning for the future</td>
<td>18</td>
</tr>
<tr>
<td>2.4 Machine learning</td>
<td>18</td>
</tr>
<tr>
<td>2.5 AI teaming and collaboration</td>
<td>20</td>
</tr>
<tr>
<td><strong>Chapter 3. Point of view: Data science</strong></td>
<td>23</td>
</tr>
<tr>
<td>3.1 Artificial vision</td>
<td>25</td>
</tr>
<tr>
<td>3.2 Natural language processing</td>
<td>27</td>
</tr>
<tr>
<td>3.3 Planning for the future</td>
<td>30</td>
</tr>
<tr>
<td>3.4 Machine learning</td>
<td>32</td>
</tr>
<tr>
<td>3.5 AI teaming and collaboration</td>
<td>34</td>
</tr>
<tr>
<td><strong>Chapter 4. Point of view: Information Technology (IT)</strong></td>
<td>37</td>
</tr>
<tr>
<td>4.1 Artificial vision</td>
<td>39</td>
</tr>
<tr>
<td>4.2 Natural language processing</td>
<td>41</td>
</tr>
<tr>
<td>4.3 Planning for the future</td>
<td>42</td>
</tr>
<tr>
<td>4.4 Machine learning</td>
<td>44</td>
</tr>
<tr>
<td>4.5 AI teaming and collaboration</td>
<td>46</td>
</tr>
<tr>
<td><strong>Chapter 5. Conclusion</strong></td>
<td>49</td>
</tr>
<tr>
<td>Related publications</td>
<td>51</td>
</tr>
<tr>
<td>IBM Redbooks</td>
<td>51</td>
</tr>
<tr>
<td>Online resources</td>
<td>51</td>
</tr>
<tr>
<td>Help from IBM</td>
<td>52</td>
</tr>
</tbody>
</table>
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.
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 Spectrum™
- IBM Spectrum Scale™
- IBM Watson®
- Power Systems™
- POWER9™
- Redbooks®
- Redpaper™

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.

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

This IBM® Redpaper™ publication will assist the line of business, data science and IT teams develop an information architecture for their enterprise AI environment. It discusses the challenges 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 AI solutions including:

- IBM Watson Machine Learning Community Edition
- IBM Watson Machine Learning Accelerator
- 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
- Planning for the future
- Machine learning
- AI teaming and collaboration

This publication targets readers from lines of business, data science teams, and information technology departments, as well as anyone interested in better understanding how to build an information architecture 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 AIX® and Linux. He holds a bachelor’s degree in Computer Science from Texas A&M University in College Station, Texas. His areas of expertise include high performance computing, systems management, cloud and cognitive computing. He has presented extensively at IBM Power Technical Conferences and is a previous 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 including Linux, open source databases, and artificial intelligence frameworks and tooling. He has presented extensively on these topics across the world, including at the IBM Systems Technical University conferences. He has been an author of multiple previous Redbooks publications.

Marcos Quezada is a Fulbright Scholar with a Masters 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 artificial intelligence, machine learning, deep learning, modern data platform, 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.

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

Ann Lund, Scott Vetter
IBM Redbooks

Freddy Alves Vaquero, Luis Armenta, Ivaylo Bozhinov, Scott Campbell, Srinivas Chitivel, 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
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:
- Stay current on recent Redbooks publications with RSS Feeds:
  http://www.redbooks.ibm.com/rss.html
Introduction

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

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

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

Enterprise artificial intelligence (AI) differs from consumer AI in that it needs an infrastructure ready to adapt to changing business needs at scale. It can start as small as a proof of concept (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 of this happens on a hybrid infrastructure supporting an information architecture that can feed data effectively into the four stages of the AI workflow while keeping the enterprise AI solution available and secure.

IBM and IBM Cognitive Systems provide industry-leading enterprise AI infrastructure for machine learning, deep learning, and inference to fuel new thinking and capabilities across your business. Your organization needs AI infused in 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 at the same time complies with traditional mission critical demands and requirements including:
- Faster time to results and accuracy
- Increased resource utilization
- Simplified management
- Enterprise grade solution

Meeting these demands and requirements with an enterprise AI infrastructure will allow your AI solutions to 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 the publication we discuss the enterprise AI workflow. This section introduces the four main stages: Ingest, Preparation, Training, and Inference, and mentions some of the challenges faced at each of the stages.

![Image](image.png)

Figure 1-1 The AI workflow

Other sources of literature on the subject may have greater or fewer stages. These stages are expansions of sub stages or compressed stages of the workflow we use in this publication. A four stage workflow is used to illustrate the examples 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 on 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, 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 partnering 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 like:

- 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**

Once the dataset is ready the important work of training is undertaken by the data science team. This is a very compute intensive stage 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 as 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 datasets.

### Inference

Once the training phase delivers the best model for the business challenge it is time to 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 quick description of the offerings available and their capabilities as well as where to go for further information.

IBM’s artificial intelligence offerings deliver a truly hybrid open source-based AI stack positioned to support your AI implementations from the POC stages to production at scale.

![IBM hybrid open source based AI stack](image)

*Figure 1-2  IBM’s hybrid open source based AI stack*
1.2.1 AI cognitive infrastructure

These are the IBM Power Systems and IBM Storage hardware offerings specifically designed to support your on premise AI workflow. They include:

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

Please refer to the following publications for more information on the hardware offerings to support your AI use.

- *Cognitive Systems Featuring the Power System AC922*, REDP-5555
- *IBM Power System AC922 Technical Overview and Introduction*, REDP-5494
- *IBM FlashSystem 9100 Product Guide*, REDP-5524

1.2.2 Organize 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 continues 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 may 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 information architecture tailored to that need.

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

**IBM Spectrum Scale**

IBM Spectrum Scale is a unified file and object software-defined storage for high performance and large scale workloads for hybrid cloud. It offers automated, policy driven, tiered storage management matching 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 HDFS to reduce 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.


**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 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 please visit: https://www.ibm.com/us-en/marketplace/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 quicker. Interactively discover, cleanse and prepare your data with a built-in data refinery.
- Curate and shape analytical assets, including data, machine-learning models and notebooks.
- Protect data misuse and confidently share assets, with automated, dynamic masking of sensitive data elements, and govern with active policy management.
- Understand data quality, data lineage and distribution through data-profile visualizations, built-in charts and statistics.

For more information please visit: https://www.ibm.com/cloud/watson-knowledge-catalog

1.2.3 Build AI

Once you have your labeled datasets ready it is time to start working through getting the best model with the highest accuracy. This task varies as data science professionals try a variety
of different methods. It is a good approach to include tools in the mix where you can share access to this stage to non data science professionals who have the knowledge of the meaningful industry sector’s business questions.

**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 a machine learning environment within minutes with a few simple commands.

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

For more information please visit: https://ibm.biz/poweraideveloper

**IBM Watson Studio**

IBM Watson Studio is a data science and machine learning platform. It helps enterprises simplify the path from proof of concept to production at scale, speeds data exploration and model development and training, and helps scale data science operations across the AI workflow. IBM Watson Studio allows organizations to tap into data assets and inject predictions into business processes and modern applications. It's suited to hybrid multicloud environments that demand mission-critical performance, security and governance — in public clouds, in private clouds, on-premises and on the desktop.

Other capabilities include:

- Automate data preparation, feature engineering, hyperparameter optimization and ensembling.
- Explore data and use machine learning with enhanced visual modeling.
- Visually program for deep learning with an intuitive drag-and-drop, no code interface in Neural Network Modeler.

For more information please visit: https://www.ibm.com/cloud/watson-studio

**IBM PowerAI Vision**

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

Other capabilities include:

- Rapid identification and labeling of datasets.
- 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 please visit: https://ibm.biz/poweraidevelopment

**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 scientists and line of business can develop trusted machine learning models. This automatic machine learning platform includes functionality for data visualization, feature engineering, model interpretability and low-latency deployment.

Other capabilities include:

- Automatic feature engineering.
- Machine learning interpretability.
- Natural language processing.
- Automatic scoring pipelines.
- Support for time series data.
- Bring your own recipes.

For more information please visit: https://www.h2o.ai/products/h2o-driverless-ai

### 1.2.4 Deploy, run, and manage AI

Once you have achieved the accuracy required for your trained model it is time to put it to work. With IBM and IBM Cognitive Systems you can deploy models on the cloud or on-premise, with a hybrid cloud infrastructure.

An enterprise AI solution needs to be cost effective, efficient, and secure, as with any mission critical application. 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 desire to increase resource utilization is a key objective for many enterprise clients that know exactly how to accomplish this for their point of sale systems or their data warehouses, but they may struggle to do this with their machine learning and deep learning tasks.

**IBM Watson Machine Learning Accelerator**

IBM Watson Machine Learning Accelerator 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 bring your own. Bigger training jobs and/or shorter completion times by distribution across multiple GPUs and/or nodes. Achieve better accuracy and insights without compromise and support for large models.
- Faster time to results and accuracy reducing the time for any single operation within the AI workflow can result in hours, days or even 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 proven support for Kerberos, SiteMinder, AD/LDAP and OS authentication.
- Authorization: Fine grained access control (RBAC), Spark version lifecycle management, notebook updates, deployments, resource plans, reporting, monitoring, log retrieval, and execution.
- Impersonation: Allow different lines of business to define production execution users.

For more information please visit: https://ibm.biz/poweraidevelopment

1.2.5 Operate AI

Once your model is in production there is a need to understand how it is performing and when it will need to go through the enterprise AI workflow all over again to maintain accuracy. Depending on the use case it may need to integrate with other tools to build a solution. An audit may require you to provide explanations for your model’s results. For these requirements it is important that your information architecture brings 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 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 please visit: https://www.ibm.com/cloud/watson-openscale

IBM Video Analytics

IBM Video Analytics turns digital video into valuable information. 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 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 please visit: https://www.ibm.com/support/knowledgecenter/en/SSKRA3_1.0.0/va/kc_welcome.html

### 1.3 How this Redpaper is organized

This publication’s goal is to promote understanding between line of business, data science and information technology professionals. The time has arrived for artificial intelligence solutions to go from proof of concept to production at scale. This can only be achieved by having tight communication between these three teams to define and create an information architecture ready to support their mixed needs and accelerate the AI workflow.

To promote this understanding we have focused 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 machine learning and deep learning techniques to create the models fueling the enterprise AI solutions that will provide answers.

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

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

1. Artificial vision
2. Natural language processing
3. Planning for the future
4. Machine learning
5. AI teaming and collaboration

These use cases were selected by the authors: selecting those generating the most interest around the world. We then brainstormed and came up with fifty questions or concerns asked in hundreds of client engagements and grouped their answers at the relevant intersections of a persona point of view and an use case.

This creation process allowed us to come up with a flexible publication that can be read in its entirety, with a a focus on a persona, or with a focus 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 have been 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 within an airport using existing security camera feeds.
- Finding faults in processor chips using high definition images from the production line.
- Automatically finding boundaries of properties for insurance valuations.

**Natural language processing**
Natural language processing uses techniques and technology to allow us to 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 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 pro-actively before a failure stops work.
- Predict and prevent fraudulent credit card transactions as they happen.

**Machine learning**
Machine learning techniques allow businesses to 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 have the ability to share their data and expertise, collaborating to create an enterprise AI solution.
- Data scientists collaborate with subject matter experts to build specific models.
- Common datasets 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 will 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 you are working to solve will influence the most appropriate tools to use.
Point of view: Line of business

Information technology (IT) has spread to all the processes within organizations. Conversely, in recent years the information technology 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 professionals have taken ownership of IT investment budgets, IT project steering committee decisions, and even the definition of what technology solutions to implement. Artificial intelligence solutions are no exception to this, and we would argue that it is where the previously described phenomenon is strongest.

Line of business professionals are concerned with finding answers to business questions, and those answers should be supported by evidence. Artificial intelligence (AI) allows them to create that evidence based on data and not static rules. Not any data, but the data 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 right people at the right time. Line of business professionals should pay close attention to their organization’s information infrastructure strategy to ensure it supports the creation of evidence at the same pace as data is produced by their company and its surrounding environment.

Artificial intelligence can help lines of business to answer the following kinds of business questions:

- How much or how many? Sales 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: line of business 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 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 to the production stage, by creating solutions based on IBM Cognitive Systems Artificial Intelligence offerings.
2.1 Artificial vision

One of the most common use cases for deep learning technology in business today is in the field of artificial vision, also known as computer vision. These techniques allow businesses to use large quantities of visual data, including images and video, to train a deep learning 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 allow. The models created can be general purpose, such as identifying people within a video, or more bespoke to a business and their visual data, and therefore solving 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 new work orders. This 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 deep learning models using their own data and their own expertise. With these models they can build a range of enterprise AI solutions to the challenges they face, reducing expenditure on visual inspection, reacting more rapidly to specific situations, or creating new 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 wastage.
- Detect aggressive or inappropriate activity and alert the appropriate authorities.
- Count the number of infected cells in a blood sample or pathology slide.
- Real time translation of sign language to improve communication.

In order to create a high quality deep learning model for artificial vision, you will need a well labelled dataset of example images or video. The quality of this labeling is crucial, as this is the data that is used to train the system. Quantity is also an important factor, as the more varied labelled examples you have the higher your likelihood of getting a high accuracy model.

PowerAI Vision includes labeling tools that allow subject matter experts to quickly and clearly label the images in a dataset. The interface allows them to categorize images, as well as label objects or actions of interest, without the need to work with a data science team or learn specialized tools. There are additional functions available to increase the speed of creating a large high quality dataset, including:

**Auto labeling** labeling new images by using a model trained on a smaller number of images, which then requires confirmation of the new labels rather than manually adding them.
Data augmentation  Creating new images by taking images that are already labelled and performing transformations, like rotation or blurring.

Using these techniques, a subject matter expert can rapidly label a dataset and train a deep learning model to help solve a real business challenge. This can then be deployed as an API service, allowing the model to be used as needed, or integrated into a specific application.

Artificial vision models trained using PowerAI Vision can also be used to identify objects in live or recorded video streams using IBM Video Analytics. This solution works with a variety of video sources and automatically identifies and tracks objects of interest, and trigger alerts or responses as needed. 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. This makes it possible to easily search video information for items of interest.

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

Artificial vision is allowing businesses to create new 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 is the set of technologies and methodologies from linguistics, computer science, information engineering, and artificial intelligence 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 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.
Many NLP solutions can be created using APIs, like Watson Natural Language Understanding, available on the IBM Cloud™. But there are use case requirements and industry standards that might require an on-premise infrastructure:

- Intellectual property control
- Privacy or confidentiality concerns
- Storage volumes required
- Processing (inference or training) time and/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 tooling is based on open source standards so your enterprise NLP solutions will be enabled to take advantage of the latest developments on NLP techniques.

If there are limited data science skills within your organization you still can answer business questions 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 don’t even know what the steps or stages it takes to build these solutions are, 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, transparency with machine learning interpretability (MLI).
- Facilitate and accelerate how these models are put into production.

If your solution only deals with understanding text then H2O Driverless AI is the best for auto machine learning. When you need to understand where that text is located in the document you can utilize IBM PowerAI Vision.

Your organization may have all the required data science skills. In this situation it is very important to start small with quick wins for the business. You can achieve this by using Watson Machine Learning Community Edition (Watson ML CE). Just as typically today users do not download the Linux kernel source plus all associated GNU utilities and compile them into a distribution that works for your business; downloading and compiling all the required frameworks for machine and deep learning is not efficient. Watson ML CE is built mostly of 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 proof of concept is ready for production the next steps involve model lifecycle management and deployment. For that you can Watson Machine Learning. When multiple teams start developing multiple machine learning or deep learning models you need to build your company’s AI cluster with Watson Machine Learning Accelerator.

All of these options will need access to the same resource: data. The cloud is great for uploading your data but will quickly stop being financially viable when your solution needs to bring that data back to your local offices, for example. There are additional challenges when you need to 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 datasets 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 line of business (LOB) perspective, incorporating artificial intelligence (AI) to help guide business choices based on intelligent predictive modeling is a key goal. Embedding into business processes predictive analytics that utilize AI for more accurate results relies on utilization of data the businesses has stored over the years.

Incorporating machine learning (ML) into analytics and decision-making can be utilized across many industries. For example:

- Insurance: detect fraudulent claims, optimize insurance quotes.
- Commercial banking: score credit risk, perform fraud-detection for credit charges.
- Government: predict fraud, optimize waste systems and traffic.
- Manufacturing: model product quality and defect detection, optimize logistics.
- Retail: customer loyalty, cross-sell/up-sell, accurate demand forecasting.
- Healthcare: medical research, predict patient condition change.
- Transportation: Optimize route planning, 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 increased number of credit products accepted per customer with more accurate cross-selling for increased revenue.

IBM has partnered with a leading automated machine learning vendor, H2O.ai, to bring their H2O Driverless AI (“a Data Scientist in a box”) product to the IBM GPU-accelerated Power Systems AC922.
2.4 Machine learning

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

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

Using machine learning techniques, businesses are able to make better use of the data they collect in order to inform decision making processes, to reduce expenditure, or to 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 machine learning models and capabilities into their offering to help customers manage the vast quantities of unstructured data they are collecting. They have added functionality based on machine learning, including tools to automatically process sales orders, to classify new data coming in, and to search for potential GDPR violations.

In order to create a successful machine learning project, you need to understand the business challenge you are trying to solve, and then have the right data and skills in place to find the required insight. The time taken to build up an appropriate solution to your challenge is affected by a number of factors, including:

- The quality and quantity of the data sources in use, and how well they fit together.
- Use of the right algorithms and techniques on that data, which is an iterative process.
- Performance of the training platform used to train and test the models 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 machine learning solutions for bespoke use cases. They can be used to accelerate the time to accuracy in a variety of ways to get value from a machine learning project faster.

Watson Machine Learning Community Edition is an easy to install software bundle that includes a number of common machine learning and deep learning frameworks and tools that are optimized for Power Systems hardware. Within this bundle are the Snap ML libraries, that provide CPU- and GPU-accelerated versions of common machine learning algorithms. This allows data scientists to train their models up to 46x faster than on other systems, allowing them to iterate more rapidly and get higher accuracy in less time. These capabilities are directly compatible with the most commonly used machine learning libraries, so data scientists 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 utilize. Watson Machine Learning Accelerator allows you to create an environment for teams to collaborate on machine learning projects, and use the resources that they need in a flexible and scalable way. This ensures efficiency, as data and resources can be shared amongst them, along with the insight and experience brought by the individuals. Watson Studio
extends this further by adding in data management, curation, and sharing across multiple projects and teams to speed innovation.

In businesses where machine learning skills are not as prevalent, then toolsets like H2O Driverless AI from H2O.ai provide automated machine learning capabilities. Within the graphical interface, it is possible to quickly visualize the data available, then create models using multiple machine learning algorithms to determine the best option for the data. This speeds the time to get a high accuracy model, allowing the data scientists to 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. This allows them to handle 90% of credit applications through these systems, offering more credit products whilst better predicting payment defaults of existing customers.

Machine learning has the ability to transform the way that businesses operate, making data driven decisions and identifying ways to reduce costs or increase income. Data science teams can use the right tools to share data and content, train models faster, and run more iterations to get higher accuracy.

### 2.5 AI teaming and collaboration

Businesses today implementing a unified artificial intelligence (AI) strategy are discovering the need to balance: the data science requirements for computing resources; realistic limitations of actual physical resources (on or off-premise); 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 be able 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 new skills, with cloud-like rapid provisioning.
- The information technology (IT) group wants to provide an environment that is secure, highly available and scalable to meet the business demands.
- The lines of business and the company executives want all of the above at a price that meets budgets and allows prioritization based on the business needs.

**Real world example:** A collaboration between two universities (medical center and imaging science institute) utilized IBM Power Systems servers and IBM Watson Machine Learning Accelerator to increase the precision with which doctors were able to identify cancerous brain tumors in under-sampled MRI images. The deep learning 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 utilize those resources can be paramount. IBM Watson Machine Learning Accelerator 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 with our Elastic Distributed Training 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 collaboratively model design, IBM Watson Studio provides a simple web-based portal for data exploration, data preparation and model development. Popular notebook technologies 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 deep learning model, IBM PowerAI Vision allows the subject matter expert from the line of business to quickly create image classification, object detection or action detection models using a simple clickable GUI with no coding required. This tool will help 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 absolutely critical for all enterprise AI solutions. Having a common repository available for your GPU-accelerated servers enables many of the advanced capabilities mentioned above. IBM Spectrum Scale is a high performance, highly available, scalable, file management software defined storage solution that supports access to petabytes of data under a single namespace reducing the need for multiple copies. Distributed deep learning requires high speed access to shared data and IBM Spectrum Scale provides that - either in a pre-configured building block (storage and management servers) or as software only, running on existing storage hardware.
Point of view: Data science

In today’s businesses, the job of the data science team is becoming ever more important. The significance of data and data science increases as businesses look to reduce costs, automate processes, or create new innovative products or services. Often the data science team is called upon to bridge departments, linking the business requirements of a project with the IT capabilities required to solve the challenge. They may be required to understand:

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

The AI offerings on Power Systems provide capabilities to reduce the burden on data engineers, data scientists, and data analysts when working on machine learning and deep learning enterprise AI solutions. These allow the data science team to concentrate on higher value tasks.

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

**Ease of use**

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

- Easy discovery and curation of appropriate business data from multiple sources.
- Tools to allow subject matter experts to label data in a quick and efficient way.
- Capabilities to rapidly iterate over multiple datasets and algorithms to discover the best combinations.
- Automate common tasks and repetitive actions like data visualization, and hyperparameter tuning.
Deploy trained models for testing and inference in automated, repeatable ways.

**Collaboration**
Work together with your team and line of business to share expertise and experiences. Use tools and functionality to allow people to work together more easily:
- Work together on datasets, notebooks, or whole projects with ease.
- Collaborate with subject matter experts to curate and prepare the best data for a project.
- Share datasets and example code as appropriate to help tackle new challenges.

**Increase 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 making use of appropriate accelerators and techniques.
- Run multiple variations of a deep learning or machine learning training run at the same time to compare performance and optimize.
- Distribute training runs across multiple accelerators or systems to increase the performance and reduce the time to accuracy.
- Optimize hyperparameters when training models through automatic evaluation and recommendations.

**Improve efficiency**
Get the most out of your limited resources by using the right tools to manage, monitor, and optimize the working environment. Improve overall performance by increasing the utilization 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 machine learning or deep learning projects you are undertaking, different tools are available to cover the focus areas above. There are some tools like PowerAI Vision that are targeted at image data and building artificial vision models, while others like Watson Machine Learning Accelerator support multiple data types and collaborative projects. The choice of the most appropriate tools will depend on the type of project being undertaken.

In this section we will go through our five use cases to understand how the data science teams within a business can benefit from having the right 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 based on IBM Cognitive Systems artificial intelligence offerings.
3.1 Artificial vision

There is a growing demand for artificial vision solutions in many businesses today. 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 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 to an industry, often with very subtle differences between categories. So off the shelf products can be unsuitable as they are trained for another business’ needs.

In order to create a high accuracy deep learning model for artificial vision you need:

- A suitably large dataset of relevant images and/or video.
- Ground truth labeling of the data to match the business need.
- A well defined deep learning network definition for visual data.
- Suitable computational power to process the training iterations.

In order to create a large set of well labelled data, most data scientists will need to work closely with subject matter experts in the industry to ensure that the images and video available are labelled correctly. This can be a time consuming process if the dataset 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 dataset in the first place, as few businesses have a rigorous storage and organization policy for image data, and so it is often dispersed across multiple systems in a variety of formats.

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

Labeling of images can also be simplified, using the tools available within PowerAI Vision. This interface can be used by subject matter experts to allow them to 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, this relieves some of the pressure on the data science team to concentrate on building the right neural networks and training and testing models. PowerAI Vision can also be used to rapidly train models with subsets of the dataset, which can then be used to label further images to add to those datasets. The subject matter experts then simply need to confirm or repair the labels that are generated, speeding up the process (Auto labeling). PowerAI Vision has built-in features to quickly increase the size of a dataset by adding new labelled examples from existing ones through transformations like rotate, blur, colorize, etc (Data Augmentation).

Once a dataset has been curated and labelled, the tools in PowerAI Vision allow you to quickly train a model for that data, using neural network definitions designed for:

- Image classification
- Object detection
- Object segmentation
- Action recognition
You can also add your own custom network designs, so that subject matter experts can train and retrain their own models based on your network designs as they adjust their datasets, to increase accuracy or eliminate bias. Alternatively, the labelled datasets can be exported in open formats, allowing them to be used in other data science tools.

Watson Machine Learning Community Edition includes the most common deep learning frameworks, which can be accessed via 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 dataset can be fixed - either by adding more data examples or correcting any mislabelled images. Iterative training helps to build a more accurate deep learning model for your use case, with regular feedback as to the performance of the model. Waiting for models to train before selection of the next step can be tedious, so reduced training times are beneficial.

For deep learning 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 datasets that deep learning training requires. The wide range of AI offerings on Power Systems are all supported on GPU-accelerated servers like the Power System AC922 model. This server supports NVIDIA Tesla V100 GPUs connected to the POWER9™ CPU using NVLink 2.0 connectivity, offering larger data bandwidth to the GPUs to maximize their performance. These benefits can be seen in the base deep learning frameworks from Watson Machine Learning Community Edition, as well as the higher level tools like PowerAI Vision. This allows you to reduce training times and therefore iterate more rapidly to help tune your dataset, hyperparameters, and model to get the best accuracy.

Another way to maximize the performance of deep learning training is to spread the work across multiple GPU devices or even multiple systems.

Watson Machine Learning Community Edition includes Distributed Deep Learning that allows data scientists to distribute a training run across multiple GPUs and multiple nodes, automating the transfer of data, network definition, and parameters between systems to share the processing and therefore producing faster results.

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

To overcome this limitation, Watson Machine Learning Community Edition includes Large Model Support. This function is available for Tensorflow, PyTorch, and IBM enhanced Caffe, and allows models to be built 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 deep learning 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. As models are trained from a dataset, they should be tested for performance and bias before modifications are made to the data, the network design, or the hyperparameters in order to train the next iteration. This iterative approach allows you to build a model that is accurate enough to solve the business challenge, without unwanted bias or error.
PowerAI Vision makes it easy to deploy a model for testing within the graphical user interface. This creates an API interface that can be used to check the model against a test dataset, held back from the training data. There is also the ability to test the model within the interface by uploading test images. This capability allows you to easily confirm that your model is providing the expected results, and identifying any problems. PowerAI Vision also provides a range of metrics for each trained model to help identify how datasets might need to be modified to provide better results.

It is also possible to deploy these trained models for long term use, allowing them to be used in a production environment or connected to a specific imaging system. The models can be deployed using IBM PowerAI Vision Inference Server, which provides the same API functionality as is seen on the training system. These deployments are containerized, and therefore provide a very flexible and scalable method for continuous inference use. Running the models on GPU-accelerated systems still offer the best performance, however 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 provides a resilient production ready environment for inference separate from the training system, and therefore having no impact on the training process for updating or creating new enterprise artificial vision solutions.

The big challenges for creating enterprise artificial vision solutions include:

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

The tools and capabilities available in various AI offerings on Power Systems can help overcome these challenges, 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 machine learning application the main challenge is collecting the right data in large enough amounts to train models, and then the infrastructure resources to run the training job. On top of that, the nature of NLP solutions makes this challenge even bigger because there are still four open problems¹:

1. Natural language understanding.
2. NLP for non popular languages; often called low-resource scenarios.
3. Large or multiple documents.
4. Datasets, 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 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 a lot of labeled data available. Most organizations want to implement NLP business solutions using their own data sources. Consequently every time a new project requirement comes down the line data science professionals have to start from the beginning of the NLP AI workflow all over again.

As mentioned before, the data science team has to create new labeled datasets each time a new NLP project starts on a new domain or even just in a new language. The current state of the art does not allow data science teams to apply existing models to adapt to new data sources. This situation, due to the iterative nature of the AI workflow, requires repeating the same tasks for each new language in order to create a new dataset to create data and subsequently algorithms:

- Find best practices from approachable research papers. Try to replicate them.
- Understand the data 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 right tooling and infrastructure plays an important role:

- Where is the needed data stored?
- What is the data subject?
- How to quickly parallelize NLP data preparation tasks?
- How to accelerate training for several projects at the same time?

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

One important requirement for data science teams working NLP AI workflows is to interact with the data in a collaborative way. This 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. IBM Spectrum Discover changes the way data science professionals interact with data, offering role based access control so different teams can confidently share data across silos. At the same time, and more critical to the NLP solutions team, is the speed with which text datasets can be created and labeled. IBM Spectrum Discover provides data

¹ For more information on NLP’s four open problems: http://ruder.io/4-biggest-open-problems-in-nlp/
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. It can also allow you 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 datasets. Your data science team needs an information architecture 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 with IBM Spectrum Scale coupled to your organization’s data lake and IBM Watson Machine Learning Accelerator to provide the infrastructure to support an AI as a Service (AIaaS) model.

IBM Watson Machine Learning Community Edition can also be considered when the project is just starting and there is no need of a shared infrastructure. 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.

Another side of data access is collaboration on data, IBM Spectrum Scale allows you to work on multiple projects across multiple sites with the same data. If collaboration is key the 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 becoming organization’s AI bottleneck, as more projects require their skills and time. There are AI projects that don’t need the expertise of a seasoned data professional. IBM Cognitive Systems offerings can help in those situations with tools to democratize access to AI. These tools can be used for automatic machine learning and deep learning development, allowing for the execution of simpler NLP projects without heavy intervention from the data science team. They are also useful tools to allow subject matter experts within the lines of business to help in the labeling stage of new domain or language text datasets. 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 line of business 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 using the most popular techniques like TFIDF, CNN, and GRU. With built-in TensorFlow, H2O Driverless AI can also process larger text blocks and build models 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. BYOR enables data science professionals to quickly leverage critical NLP community knowledge by incorporating customizations and extensions to the platform. These customizations are Python code uploaded at runtime including:

- Custom machine learning models
- Custom scorers (classification or regression)
- Custom transformers
- Custom datasets

IBM PowerAI Vision

When your NLP data science teams have limited skills they can use other sources of data to create their text datasets including print media, like digital newspaper articles, magazines, blogs, and others. Images require them to apply deep learning 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 deep learning 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 utilize AI to provide the business with positive results they have to 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/or local sources and split the data appropriately between training and testing subsets.
- Train and deploy the model(s): using the data feature engineering, model tuning, model selection.
- Verify the results: review accuracy, visualizations can be useful to help understand and provide interpretability (explanations) of the results.

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

Industries in which you can incorporate machine learning (ML) into analytics and decision-making are varied and can include insurance, energy/utilities, banking, government, manufacturing, retail, healthcare, transportation.
IBM has partnered with a leading automated machine learning vendor, H2O.ai, to bring their H2O Driverless AI product designed 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 respectively. It can also be used by subject domain users or line of business professionals to quickly provide results with no coding required.

- Datasets can be ingested from local files (many formats supported) or via remote data connectors for HDFS, Amazon S3, Google Cloud Store and many others.
- Experiments are where H2O Driverless AI analyses the data based on the selected target variable for results and automatically selects ML models based on included feature engineering recipes pre-built by expert data scientists. Expert level customizations are selectable if wanted.
- There are three main knobs that can be utilized to adjust the outcome of an experiment. Adjusting one knob does affect the others:
  - Accuracy
  - Time
  - Interpretability
- Machine learning interpretability provides human-level understanding of the math and processes behind the machine learning model and its results.
- In addition to the graphical user interface as the primary avenue 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 Machine Learning Community Edition (Watson ML CE) for this purpose. Watson ML CE provides pre-compiled (both ppc64le and x86_64) versions of popular machine and deep learning frameworks. Watson ML CE was co-developed with the IBM GPU-enabled POWER9 processor-based server (IBM Power System AC922), to take advantage of the advanced hardware capabilities. For more information on the Power AC922, please see the IBM Redbook publications Power System AC922 Technical Overview and Introduction, REDP-5494 or Cognitive Computing Featuring the Power System AC922, REDP-5555.

In addition, IBM Research developed an optimized, enhanced drop-in library with accelerated machine learning libraries to speed processing on CPU and GPU in the Power AC922 called Snap ML. For more information, see 3.4, “Machine learning” on page 32.

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 the following locations:
https://www.h2o.ai/try-driverless-ai/

Both Watson ML CE and H2O Driverless AI are available as installable software packages 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.

Documentation and additional information for Watson ML CE can be found at the following locations:
Documentation and additional information for H2O Driverless AI can be found at the following locations:

- [http://docs.h2o.ai/driverless-ai/latest-stable/docs/userguide/index.html](http://docs.h2o.ai/driverless-ai/latest-stable/docs/userguide/index.html)
- [https://github.com/h2oai/tutorials/tree/master/DriverlessAI](https://github.com/h2oai/tutorials/tree/master/DriverlessAI)
- [https://www.h2o.ai/try-driverless-ai](https://www.h2o.ai/try-driverless-ai)

### 3.4 Machine learning

As machine learning (ML) becomes more prevalent, there is increasing pressure on data science teams to build new solutions and capabilities using data driven machine learning techniques. Along with the increase in data volumes, the requests from businesses for new innovation and new ML models is increasing. In order to meet growing demand, data science teams need to overcome a number of challenges.

Data collection and preparation is a key first step in any data science project, including machine learning projects. It is important to identify the right sources of data to help solve the business challenge. That data then needs to be organized and curated in a way that allows disjointed data to 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, searching by file type, metadata entry, or custom tags. This makes it possible to pull out all of the files within a business that have been labelled as a certain type, or hold a particular type of data. This tooling also allows you to automatically add custom tags to files as they are created or processed, for example allowing you to keep track of project usage for governance. These tools make it easier to collate the data required for a machine learning project by easily finding the data most relevant to the business challenge.

There are benefits to be gained from collaborating on projects, to share 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 datasets and connections to external data sources, as well as collaborating on model building through notebooks, IDEs, and visual modelling tools. Data scientists can contribute to multiple projects, with datasets and other assets managed within the projects to share or limit access to other teams as needed. There is notebook support for Python, R, and Scala, as well as RStudio support and tooling for creating model flows using visual interfaces.

Like other data science work, machine learning projects generally require a highly iterative approach. Data scientists work with a range of different machine learning algorithms, and work with the features available in the data to derive the greatest insight. This may involve feature engineering approaches, as well as identifying other sources of data to add to the working set. Each new combination of data and algorithm requires training of a new model, which takes time to complete, so as you increase the number of options you work with, the time requirements of the project also go up.
In order to minimize the impact of multiple iterations across the project, you can reduce the time required to train each model. Watson Machine Learning Community Edition includes the Snap ML libraries. Snap ML adds GPU acceleration to some of the most common machine learning algorithms, speeding up training times up to 46x\(^2\). Snap ML introduces parallelization to the algorithms, allowing specialized accelerators like GPUs to improve performance, but also to improve scalability over devices or systems for larger datasets. 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 machine learning code with minimal changes. The algorithms 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, however if you use the Snap ML Python API, then any functions 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 using 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 accuracy model. Using Watson Machine Learning Accelerator allows you to 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 assists in finding the most suitable algorithm for your dataset. This tooling automates a lot of the processes used 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 datasets, to help identify outliers or anomalies, or see where features are heavily correlated. This can help a data scientist to better understand the dataset, 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 machine learning algorithms into business processes is that of trust. Traditional methods of building or automating business processes have been based around implementing rules, which creates processes that are easy to explain. This allows people to interpret how a decision has been made by following the rules that implement that choice. With machine learning techniques, business processes can be built using data driven techniques which create effective models, but cannot easily be explained.

H2O Driverless AI also includes interpretability functions, which can be used to take a complex machine learning model, and explain how it makes decisions, and which features have the largest impact on the result. It uses machine learning techniques to show more easily understood linear models or decision trees in place of complex models. These can be used to explain how decisions have been made in individual cases, which can be shared with line of business professionals or used to confirm that no unwanted biases exists.

\(^2\) Further information on Snap ML can be found at https://www.zurich.ibm.com/snapml/.
The final challenge in a machine learning project is understanding how best to deploy a trained model for business use. This should follow the iterative training process including a period of testing to ensure the suitability of the model performance and no unwanted bias or unfairness exist. At this stage, the model needs to 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 hugely valuable. Managing the model lifecycle is also of value, to track different versions of models as they are retrained on new data coming in.

Watson Machine Learning Accelerator has model management and deployment functionality. With 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 functionality to test the models that you have created quickly and easily. H2O Driverless AI also includes capabilities to export trained models for inference with runtimes for Python and Java that can be incorporated into applications.

Throughout the development of a machine learning solution, there are techniques that can be employed to get to higher levels of accuracy faster, form better understanding of the data, enjoy faster training iterations, and simplified model management, testing, and deployment.

### 3.5 AI teaming and collaboration

Data science is all about teaming and collaboration to produce business results from enterprise AI solutions. The number of AI projects in organizations continue to proliferate beyond the capabilities. 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.

Given the growth in the number of data science professionals, the ability to work together in a coordinated fashion and/or 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 server(s) or cloud instances, having a unified workbench where tools, data, and models can be shared allows for re-use, sharing, and consistency.
IBM has an end-to-end AI strategy with tools for data management, model management and ongoing production support.

![IBM hybrid open source based AI stack](image)

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

IBM Waston Studio provides a web-based developer tool bench 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 to work with:

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

Watson Machine Learning works in conjunction with Waston Studio to provide the training, execution and deployment of the actual machine learning and deep learning models. The primary interaction with Watson Studio and Watson Machine Learning is via a graphical, web-based UI, but command-line and REST APIs are available for programmatic access.

IBM Watson Machine Learning Accelerator is a Spark-based, multi-tenant offering that is designed to speed up machine learning and deep learning development on a cluster of GPU-enabled servers. The deep learning tasks can be self-contained or come from Watson Machine Learning or Watson Studio. Watson Machine Learning Accelerator (WMLA) has many advanced features of interest to the data scientist looking to increase their efficiency:

- Distributed deep learning and GPU scheduling to fully utilize the limited GPU resources across the cluster. This includes:
  - Prioritization based on administratively-defined policies between groups and/or users.
  - Preemption of GPU resources (if required) based on policy.
Dynamic addition or removal of GPU resources from running training jobs using the Elastic Distributed Training feature based on priorities and/or the fairshare resource allocation algorithm. This capability allows data science professionals to be unaware of the hardware foundation. 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, underfitting, etc. 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 deep learning frameworks that come with Watson Machine Learning Community Edition like TensorFlow, PyTorch, Caffe, Keras, Bazel, ONNX, NVIDIA Rapids cuDF/cuML and more. There is also a capability to bring your own framework and extend the capabilities of the tools.

- Data ingest, preparation, and transformation tools based on Apache Spark including splitting of datasets into training, validation and testing subsets.

**Real world example:** Large US banks utilize Corporate Model Risk departments to monitor and maintain interpretability in machine learning/deep learning 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 as well as 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 resource. Using the IBM Power System AC922 with it's fast CPU-GPU NVLink connection (moving large data through faster) and Watson Machine Learning Accelerator for efficient scheduling of the GPUs allow them to achieve higher accuracies faster.

IBM Watson OpenScale focuses on the production operation of ML/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 used in decision making and the weight of each attribute.

The Watson family of products mentioned in this section are available both in an on-premise Kubernetes-based cloud or off-premise cloud. Multiple vendors clouds are supported. A free trial as well as additional information can be found at the following locations:

- [https://ibm.biz/poweraideveloper](https://ibm.biz/poweraideveloper)
- [https://www.ibm.com/cloud/watson-openscale](https://www.ibm.com/cloud/watson-openscale)
Chapter 4. Point of view: Information Technology (IT)

The use of artificial intelligence (AI) has spread throughout organizations today in both large and small projects. Historical implementations of computing power have been under the management of IT departments. Today however, in a race to quickly implement enterprise AI solutions used by data science professionals, the various lines of business (LOB) have bypassed 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 lines of business have created their own miniature IT organizations, sometimes called “shadow IT”.

This shift has 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 already utilized 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 defined by IT, so exceptions must be requested, additional skills acquired, etc.
- Cloud-based resources that the IT organization need to support add another conflict with policies, procedures, security, and hardening of those cloud assets may not be as easy as the on-premise resources managed by IT.

For many new workloads, information technology 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 who aren’t continue to lose the budgets associated with the support of the infrastructure needed for these projects. That money now goes to shadow IT projects, that don’t benefit from the oversight or experience of the IT department. In order to bridge this gap, IT departments need to adapt their services and offerings to meet the new demands of the business.

The cycle of computing swings like a pendulum between strict management and consolidation to untethered distribution of compute resources. Companies typically have
compute resources in both places: lines of business 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 lines of business.

For some companies this may mean pushing everything into the cloud with a prescribed list of cloud vendors and a consolidated cost structure. For other companies, it may mean a consolidation of data centers with greater use of virtualization and/or private clouds to increase utilization and flexibility of compute resource. As before, the end goal may 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 implementation of AI initiatives with the lines of business and data science teams:

- Where does the data for AI model training reside with respect to the GPU-enabled compute resource 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? The use of hybrid cloud is a key method for implementation to get the best of both worlds (on- and off-premise). This item often involves determining the best infrastructure for your AI workloads within the available budget.

- For training, how can I best utilize the limited GPU-enabled compute resource 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 said data?

Note that these items listed above are all in support of the business objectives, not instead of the business objectives.

In summary, the IT organization is there to support the line of business and the data science teams to produce the very 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 information technology department.

In this section we will go through our five use cases to understand how IT can address the business requirements for implementation of AI initiatives by creating solutions based on IBM Cognitive Systems artificial intelligence offerings.
4.1 Artificial vision

Advances in artificial vision technologies are creating new challenges for IT departments, as different lines of business look to implement these capabilities. This means that data science teams require access to the latest tools, frameworks, and libraries, whilst also wanting a more flexible service allowing them to test multiple approaches with different parameters. It is then up to the IT teams to provide these technologies rapidly, whilst also focusing on key factors such as:

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

In order to balance these factors with the requirements of the data science teams for new tools and functionality, IT departments may require new deployment methods or management techniques. A common approach is to introduce cloud-like capabilities, allowing data scientists to quickly create new 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 a number of steps that data science teams need to iterate through to build a high accuracy model. This involves:

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

Each step has different requirements of the underlying technology and the way it is configured, and in some cases multiple systems or environments will be used concurrently. The PowerAI Vision toolset includes tools to accelerate and simplify each step of this process, providing for rapid development of a variety of deep learning 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 methods include:

*Stand-alone system installation*

PowerAI Vision needs to run on GPU-accelerated systems like the Power Systems AC922 server. It is possible to install the software on a single system, making use of all of the resources available to run every aspect of the model creation process. Model training or deployment can use GPU resource as needed as long as it is 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 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 required to deploy a working instance of the software stack. This can then
be deployed through private cloud management systems like IBM Cloud Private or Red Hat OpenShift, and allows multiple instances of PowerAI Vision to run on a single cluster whilst making use of all available resources including GPUs. This method also allows other workloads to 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 datasets and trained models. It is also possible to import image data from a variety of different sources, either directly through the simple web based interface, or through the REST API built into 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.

The use of a containerized approach makes it easier to deploy upgrades and/or perform maintenance tasks during the life of the software deployments. To gain new features and capabilities rapidly, updated container images can be added to the system(s) and the service restarted. In a private cloud environment, all running containers can be migrated to other systems if needed for server maintenance.

The training aspect of deep learning 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, as they can be restarted if needed in the event of a failure. Data resiliency of the labelled datasets is more important, and so within PowerAI Vision it is possible to export labelled datasets 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. This is based on the nature of the deep learning model, and how the importance to the business can require a more resilient approach. The PowerAI Vision Inference Server allows trained models to be deployed in a separate environment from the training environment. This is also container based, creating and running new container instances with inference services exposed as APIs ready for applications to make use of them. These container deployments are scalable, allowing you to 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 available. The PowerAI Vision Inference Server deployments can make use of GPU accelerators where available, but can also run as CPU only containers - although at a reduced performance.

For data science teams using other tools to build artificial vision capabilities, it is also possible to run the common deep learning frameworks and tools in both bare-metal and container based installations. Watson Machine Learning Community Edition can be installed on a single system, or a cluster of systems with GPU accelerators. This can be:

**Single system installation**

Provides access to the common deep learning frameworks like Tensorflow, PyTorch, Caffe, and Keras on a single system, through the command line, Jupyter notebooks, or IDEs. Jobs can be run on all of the resources within the single system.

**Clustered installation**

Multiple systems can be used to distribute training workloads across a number of servers. Distributed Deep Learning functionality is used to spread the model of a training job across any number of GPU accelerators within the clustered environment using a shared file system for simultaneous data accessibility. This allows for faster training times as parallel tasks are shared across the resources available.
Private cloud installation

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

For collaborative environments, Watson Machine Learning Accelerator simplifies some of the common tasks of creating datasets and training models, whilst still providing a high level of control over the process to the data scientists. This can be deployed on any size cluster of systems with GPU accelerators, and uses Elastic Distributed Training to maximise the use of the resources available across all submitted training jobs. This software also manages user and role based access controls in a multi-user environments, allowing limits to be placed on users or projects, as well as controls placed on data access.

PowerAI Vision and Watson Machine Learning Accelerator both include support from IBM. When run on the Power Systems AC922 server with NVIDIA Tesla V100 GPU accelerators, and using 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 provides greater confidence in the offerings, and help with your deployment should you need it. There is also an optional support offering for the Watson Machine Learning Community Edition software bundle available if required, although the standard offering is a no charge product with community support only.

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

4.2 Natural language processing

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

Business solutions based on these technologies and methodologies include document classifiers, cognitive OCR, voice-of-customer, advertisement optimization and virtual assistants. Please see 2.2, “Natural language processing” on page 16 for a brief description of each business solution.

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

IT departments can support NLP solutions at scale by making the organizations’ information readily available to start creating training datasets. To support the more iterative nature of an NLP AI workflow, characterized in 3.2, “Natural language processing” on page 27, IT has to create an information architecture 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 information architecture 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 they need by understanding what is contained in your data lake, where it comes from, and how accurate it is. On top of that the IT teams need to support enterprise requirements like:

- Role based access control so departments can securely share data across silos as appropriate, enabling NLP data science teams to discover and access the data they need.
- Governance to comply with access restrictions so the right people have access to the right data at the right time without fearing a breach of compliance security.

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

Data science professionals dealing with NLP solutions need to parallel process stemming, counting, creating frequency vectors, and other tasks over and over again on each domain or language dataset. 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. At the same time it can minimize movement 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 have to collaborate with the data science teams to decide which solution will provide the most suitable capabilities based on the data science skills 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. These tools can be used for auto ML and auto DL allowing 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 at the line of business 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 datasets.

Finally, as more data science projects are started across a business, the IT team may need to implement a reference architecture that will allow their organization to offer a machine learning as a service environment. As data volumes grow into the petabyte range, and there are multiple teams and multiple AI projects running at the same time, tooling like IBM Watson Machine Learning Accelerator is recommended. This 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 IBM Watson Machine Learning Community Edition can be considered. 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 that businesses are using machine learning and deep learning 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. Machine learning can add accuracy to these traditionally data and compute intensive tasks. The typical AI workflow involves:

- Data ingest and formatting: selecting the proper data to analyze from local, shared and/or remote sources. Modification of the data for proper utilization by the tools might require
multiple copies if in-line modifications are not possible. Typically this requires a lot of 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 desired results. This process often streams the data through the models 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 above is the goal of the IT organization. This includes the software tooling and hardware required to meet the business goals in a timely manner.

Every organization runs on data (whether or not they realize it). Providing correct and relevant data as input for an ML analytical pipeline is the key to getting accurate results.

IBM Spectrum Scale is a high-performance, shared file system with advanced capabilities including integration of traditional Posix-file, object, CIFS, and NFS into one global namespace. IBM Spectrum Scale is utilized 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 has been designed to meet the needs of the largest data stores, and that same technology can be implemented at a customer’s location. Most importantly, this can be integrated with a customer’s existing storage if needed, or use a building block approach of software defined storage components with advanced reliability and availability capabilities to grow a new data hub.

Utilizing the correct data often means knowing information (i.e. data) about your data. IBM Spectrum Discover connects to file and object storage (both on-premise 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 selecting the proper data for predictive analytics.

IBM’s partnership with leading machine learning vendor, H2O.ai, has brought their flagship automated ML product, H2O Driverless AI to the IBM Power Systems platform. Junior data scientists and/or subject domain experts can utilize H2O Driverless AI with no coding required to achieve high predictive accuracy similar to results achieved by advanced data scientists. H2O Driverless AI utilizes automated model selection and feature engineering based on the data available. Some of the algorithms include linear models, neural nets, clustering and dimensionality reduction models and many traditional approaches such as one-hot encoding. More experienced data scientists and analysts can adjust details to fine-tune the results.

H2O Driverless AI utilizes both GPUs and/or CPUs depending on the model and available hardware resources. When utilized on IBM Power System AC922 servers, H2O Driverless AI can run on bare metal with NVIDIA CUDA 10.0, or later, drivers 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 using the nvidia-docker2 version of the NVIDIA Container Toolkit. Customers already utilizing IBM Watson Machine Learning Accelerator for efficient utilization of their IBM Power System AC922 cluster can integrate and launch H2O Driverless AI instances using the Watson ML Accelerator scheduler.

For customers who develop their own ML models with popular frameworks like TensorFlow, PyTorch, NVIDIA cuML, and more, IBM Watson Machine Learning Community Edition (Watson ML CE) is provided. Watson ML CE is available via a public Conda channel for use on a per-user basis running bare metal 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 a variety of 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: https://hub.docker.com/r/ibmcom/powerai/

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

Some of the frameworks can take advantage of an advanced distributed mode providing near-linear scaling of training times across a cluster of up to four IBM GPU-enabled servers called Distributed Deep Learning (DDL). DDL takes advantage of low-level RDMA communications possible with InfiniBand networks using traditional HPC protocols provided by IBM Spectrum™ MPI (included). Larger clusters (> 4 nodes) are possible with IBM Watson Machine Learning Accelerator in addition to Watson ML CE. Customers can use DDL with traditional Ethernet technologies, but the lower overall bandwidth and higher overhead with the full TCP/IP stack will not provide as great a performance boost.

For more information on the IBM Power System AC922, please the IBM Redbook publications Power System AC922 Technical Overview and Introduction, REDP-5494 or <working title Cognitive and the AC922>, REDP-5555.

H2O Driverless AI is available from IBM and H2O.ai. A trial can be acquired at:
- https://www.h2o.ai/try-driverless-ai/

Documentation and additional information for Watson ML CE can be found at:
- https://ibm.biz/poweraideveloper

Documentation and additional information for Driverless AI can be found at:
- https://github.com/h2oai/tutorials/tree/master/DriverlessAI
- https://www.h2o.ai/try-driverless-ai

### 4.4 Machine learning

As machine learning 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 on IT infrastructure and administration provide different challenges to those of traditional IT. Data is at the heart of machine learning, and so management of data access becomes an important consideration, but so too is user access as you manage multiple projects across your infrastructure. For a multi-user machine learning environment you need to consider:
- User management and integration to authentication.
- Data management and access control lists.
- How the users will access the resources.
- What tools and applications will be available.
- How to ensure most efficient use of resources.
- Collaboration requirements of the data science team.
For machine learning workloads, the requirements of the data science team may be quite diverse. There are a range of common frameworks and libraries that may be required, and with a rapid pace of development there may be requests for different versions. The majority of work is carried out in a Python environment of some sort, although access may be requested through the command line, web-based notebooks, or IDEs. There may be dozens of different variations of the environments requested.

Watson Machine Learning Community Edition makes the process of creating and managing environments easier, by using the Anaconda Python package and environment manager. This allows users to create their own Python environments, choosing 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 tooling pulls in all of the needed dependencies, and allows multiple environments to exist independent of each other, so multiple versions can be in use at the same time.

Watson Machine Learning Community Edition also includes Snap ML, a set of GPU-accelerated libraries for common machine learning algorithms. When running on GPU-accelerated servers like the IBM Power System AC922 this can provide accelerated training times by running threads in parallel to make use of GPU capabilities. As Snap ML includes an API service that is compatible with the commonly used SciKit Learn library, data scientists can make use of this acceleration with minimal changes to their code. This increases the speed of training, and therefore the overall performance of creating a machine learning model.

These capabilities are also available in Watson Machine Learning Accelerator, a multi-user environment for data science teams. This can be run across multiple systems, and 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, Watson Machine Learning Accelerator can integrate into your existing systems whilst offering a highly flexible data science platform. This tooling can also increase the efficiency of resources as it acts as a scheduler for deep learning and machine learning workloads. Users can access the environment through a variety of 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 is done elastically, allowing new jobs to be run interactively while others continue training. Users can choose which environments and packages they work with, offering flexibility in how they work, with GPU acceleration available for the workloads that will benefit from it.

Watson Machine Learning Accelerator also has a strong focus on systems management and administration, collecting and presenting metrics for system performance, utilization, and health. There is also the ability to record user or group usage of the resources, allowing for capping of some teams, or 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. This provides a data science workbench, enabling project teams to share access to datasets, notebooks, and other tools to collaborate on their work. Any training jobs that are created from this work can be deployed onto a Watson Machine Learning Accelerator cluster to make the most efficient use of resources and ensure everything is managed and monitored effectively. The collaboration tools allow data science teams to share their expertise, as well as their data and code, leading to improved results. Watson Studio also includes tooling 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 machine learning models is H2O Driverless AI from H2O. This provides automation of many stages of the AI workflow, reducing the burden on a data scientist, and allowing for 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 Watson Machine Learning Accelerator, sharing resources as needed with other workloads. It is also possible to install it on single systems as an independent workload. As H2O Driverless AI makes use of GPU acceleration systems like the Power AC922 server with NVIDIA Tesla V100 GPUs are 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.

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

4.5 AI teaming and collaboration

Growth in the area of artificial intelligence projects within organizations has often occurred in an unplanned fashion with separate projects cropping up all over. Data science professionals working on these projects have not necessarily communicated with each other, utilizing their own individual or departmental computing and/or storage resources. IT departments are challenged to provide a superior service to those data science teams and line of business that:

- Securely provide access to the data required to train machine learning and deep learning 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 deep learning tools, libraries and associated programs that are familiar to the data scientists and meet or exceed performance requirements.
- And do it all with cloud-like efficiency at a reasonable cost.
  - Rapid availability.
  - Simple, centralized management.
  - Flexibility to respond to surges in demand.
- Provide an information architecture to broaden the access of line of business professionals to AI tooling that can involve them in the AI workflow (IBM PowerAI Vision and H2O Driverless AI).

New applications are often built to utilize cloud-like object storage, whereas historical data in enterprises typically resides in traditional file-based repositories. IBM Spectrum Scale is a software-defined storage solution that can utilize 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 release of this publication): the Summit supercomputer at Oak Ridge National Laboratory with 250 PB of storage that can sustain 2.2 TB/sec of sequential throughput. Encryption, parallel backups, quality-of-service, and full auditing support are just a few of the features available in IBM Spectrum Scale that can support not only a large amount of data, but a extensive number of data science projects.

IBM Watson Studio and IBM Watson Machine Learning provide a web-based data science tool bench that provide data preparation tools, data visualization tools, model building,
management, and more. From these products, GPU-accelerated training jobs can be sent to IBM Watson Machine Learning Accelerator running on a cluster of IBM Power System 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 bare metal cluster for IBM Watson Machine Learning Accelerator.

IBM Watson Machine Learning Accelerator is built upon IBM Spectrum Conductor, a Spark-based, multi-tenant cluster offering that provides tools for data ingestion and manipulation, and Deep Learning Impact that supports deep learning training and inference with advanced features. Installation of IBM Watson Machine Learning Accelerator 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 as that is completely handled by the Watson ML Accelerator installer.

Watson ML Accelerator has two features that are designed to utilize multiple GPUs (on the same node or across more than one node) to reduce the overall training time and reduce latency once in the inferencing stage: Distributed Deep Learning and Elastic Distributed Training.

- Distributed Deep Learning (DDL) utilizes a popular high performance computing message passing protocol called MPI (message passing interface). IBM Spectrum MPI (included with Watson ML Accelerator) is designed to provide reliable messaging, typically 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 Gb/s or 100 Gb/s are recommended. With Watson ML Accelerator, all the Power AC922 nodes in the cluster can participate providing near linear scaling in training times.

- Elastic Distributed Training (EDT) also utilizes the network for communication, but relies heavily on the shared file system for sub task 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 comprised of Flash or NVMe with tier 1 based on SSD technology and tier 2 based on spinning disks. EDT can dynamically add or remove GPU sub tasks from the training job based on queue workload and administratively-defined priorities between users and groups.

Watson ML Accelerator provides access to many of the latest machine learning and deep learning frameworks for the data science professionals to use from the embedded IBM Watson Machine Learning Community Edition. These include current releases of TensorFlow, PyTorch, Caffe, Keras, Bazel, ONNX, NVIDIA Rapids cuDF/cuML and more. Watson ML CE also includes accelerated machine learning libraries for the Power AC922 called Snap ML.

Most importantly for IT and the data science teams, these frameworks, along with the advanced deep learning capabilities of Watson ML Accelerator (like the Spark configuration, Elastic Distributed Training, hyperparameter optimization), are all fully supported by IBM. If a data scientist desires, the Spark framework of Watson ML Accelerator can be used to add additional frameworks not include with Watson ML CE.

Additional information can be found at the following locations:

- [https://ibm.biz/poweraideveloper](https://ibm.biz/poweraideveloper)
Conclusion

This publication helped the line of business, data science and information technology professionals find common ground and work as a team in building an enterprise AI solution.

These professionals provide the following:

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

The following table shows a list of IBM Cognitive Systems offerings and capabilities at the intersection of a persona and a popular use case.

<table>
<thead>
<tr>
<th>Persona / Use case</th>
<th>Line of business</th>
<th>Data science</th>
<th>Information technology</th>
</tr>
</thead>
<tbody>
<tr>
<td>Artificial vision</td>
<td>IBM PowerAI Vision</td>
<td>IBM Spectrum Discover IBM PowerAI Vision Large Model Support</td>
<td>IBM PowerAI Vision IBM Watson® ML CE IBM Watson ML Accelerator</td>
</tr>
<tr>
<td>Natural language processing</td>
<td>IBM Spectrum Discover</td>
<td>IBM Spectrum Scale™ IBM Spectrum Discovery Snap ML IBM Watson ML Accelerator H2O Driverless AI</td>
<td>IBM Spectrum Scale IBM Spectrum Discovery IBM Watson ML Accelerator</td>
</tr>
<tr>
<td>Planning for the future</td>
<td>H2O Driverless AI</td>
<td>H2O Driverless AI IBM Watson ML CE Snap ML</td>
<td>H2O Driverless AI IBM Watson ML CE Snap ML</td>
</tr>
</tbody>
</table>
Artificial intelligence is different in many ways to the information technology solutions previously implemented in organizations. It requires a new vocabulary, new skills, and new ways of working in order 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.

As you finish this publication, line of business, data science and IT teams will be able to sit around the table and have meaningful conversations using a common language and set of tools, to discover together how enterprise AI can be put to work to propel your business with competitive advantage and create new sources of revenue.

<table>
<thead>
<tr>
<th>Persona / Use case</th>
<th>Line of business</th>
<th>Data science</th>
<th>information technology</th>
</tr>
</thead>
<tbody>
<tr>
<td>Machine learning</td>
<td>Watson Studio</td>
<td>IBM Spectrum</td>
<td>IBM Watson ML CE</td>
</tr>
<tr>
<td></td>
<td>Watson ML CE</td>
<td>Discover</td>
<td>IBM Watson ML</td>
</tr>
<tr>
<td></td>
<td>H2O Driverless AI</td>
<td>IBM Watson ML CE</td>
<td>IBM Watson ML Accelerator</td>
</tr>
<tr>
<td></td>
<td></td>
<td>IBM Watson ML</td>
<td>Snap ML</td>
</tr>
<tr>
<td>AI teaming and collaboration</td>
<td>IBM Watson Studio</td>
<td>IBM Watson Studio</td>
<td>IBM Spectrum Scale</td>
</tr>
<tr>
<td></td>
<td>IBM PowerAI Vision</td>
<td>IBM Watson ML</td>
<td>IBM Watson Studio</td>
</tr>
<tr>
<td></td>
<td>IBM Watson ML Accelerator</td>
<td>IBM Watson OpenScale</td>
<td>IBM Watson Studio</td>
</tr>
<tr>
<td></td>
<td>IBM Spectrum Scale</td>
<td></td>
<td>IBM Watson ML Accelerator</td>
</tr>
</tbody>
</table>
Related publications

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

IBM Redbooks

The following IBM Redbooks publications provide additional information about the topic in this document. Note that some publications referenced in this list might be available in softcopy only.

- *Cognitive Systems Featuring the Power System AC922*, REDP-5555
- *IBM Power System AC922 Technical Overview and Introduction*, REDP-5494
- *BM FlashSystem A9000 and A9000R Architecture and Implementation (Version 12.3.2)*, SG24-8345

You can search for, view, download or order these documents and other Redbooks, Redpapers, Web Docs, draft and additional materials, at the following website:

ibm.com/redbooks

Online resources

These websites are also relevant as further information sources:

- IBM Spectrum Scale
- IBM Spectrum Discover
- IBM Watson Knowledge Catalog
  https://www.ibm.com/cloud/watson-knowledge-catalog
- IBM Cognitive Systems developer portal
  https://ibm.biz/poweraideveloper
- IBM Watson Studio
  https://www.ibm.com/cloud/watson-studio
- H2O Driverless AI
  https://www.h2o.ai/products/h2o-driverless-ai
- IBM Watson OpenScale
  https://www.ibm.com/cloud/watson-openscale
Help from IBM

IBM Support and downloads
ibm.com/support

IBM Global Services
ibm.com/services