AI Use Cases, Infrastructure, Market Survey, and IT Reference Architectures

Trademarks and Permissions

 and other Futurewei trademarks are trademarks of Futurewei Technologies. Huawei trademarks are trademarks of Huawei Technologies Co., Ltd. All other trademarks and trade names mentioned in this document are the property of their respective holders.

Notice

The purchased products, services and features are stipulated by the contract made between Huawei and the customer. All or part of the products, services and features described in this document may not be within the purchase scope or the usage scope. Unless otherwise specified in the contract, all statements, information, and recommendations in this document are provided "AS IS" without warranties, guarantees or representations of any kind, either express or implied.

The information in this document is subject to change without notice. Every effort has been made in the preparation of this document to ensure accuracy of the contents, but all statements, information, and recommendations in this document do not constitute a warranty of any kind, express or implied.

FUTUREWEI TECHNOLOGIES, INC.

Boston Research Center

Address: 111 Speen Street, Suite 114
Framingham, MA 01701
United States of America

Website: http://www.futurewei.com/

AI Use Cases, Infrastructure, Market Survey, and IT Reference Architectures

Abstract: As the world entered the second decade of the 21st century, AI applications are changing the world rapidly. At the same time, these applications are demanding more from the infrastructure supporting them. While vendors are working on providing more powerful machines, engineers and researchers are developing innovative products and algorithms that could use more resources. The infrastructure needs to be scalable, reliable, and fast to support more computation and IO requests. As the use cases and AI algorithms being used can change frequently, an AI infrastructure must also be efficient for different algorithms. At last, as GPUs and fast CPUs are expensive, an AI infrastructure needs to be cost-effective. It must find a balance between the time to finish a job and the cost spend with it. AI infrastructure also needs to be efficient in a way how GPUs (more expensive) can be utilized to their full potentials.

In this white paper, we survey different aspects of an AI ecosystem and look at current AI market leaders in enterprise applications and different industries. Then, we gave an overview of the market opportunities of IT vendors for AI infrastructure in these market sectors. Then based on these use cases, a common reference architecture is proposed.

The purpose of this document is to discuss a set of interesting topics related to AI infrastructure, so the reader can obtain an overall picture of AI infrastructure use cases on the market. The document is having the following audiences in mind:

- *IT executives who are interested in AI use cases and AI solutions.*
- *Solution Architects who are interested in AI use cases and requirements.*
- *Sales and marketing professionals who are interested in current market conditions.*

CONTENTS

1 AI ECOSYSTEM AND MARKET TREND

1.1 GENERAL TRENDS

The AI ecosystem development is driven by both industrial companies and academic researchers. As deep learning algorithms found practical usage in our daily life, they are widely adopted in computer vision, content recommendation, autonomous driving, and life science research. AI has become one of the fastest-growing areas in the computer science field.

While big data technologies are having a wide adoption in the field of data analytics, machine learning (deep learning in particular), joined the toolbox to provide better utilization of the large amount of data collected. The abundance of training data and more advanced algorithms together proved that AI could do much more than what people expected 10 years ago. Many companies that used to invest heavily in HPC platforms are now adding AI deep learning platforms into their IT infrastructure.

On the vendor and service provider side, all tech companies are investing heavily in the AI market. Many companies are investing in the AI technology stack from top to bottom. From the lower layer components such as AI chips, chip drivers, math libraries, deep learning frameworks, job scheduling platforms, all the way to AI applications for different industries.

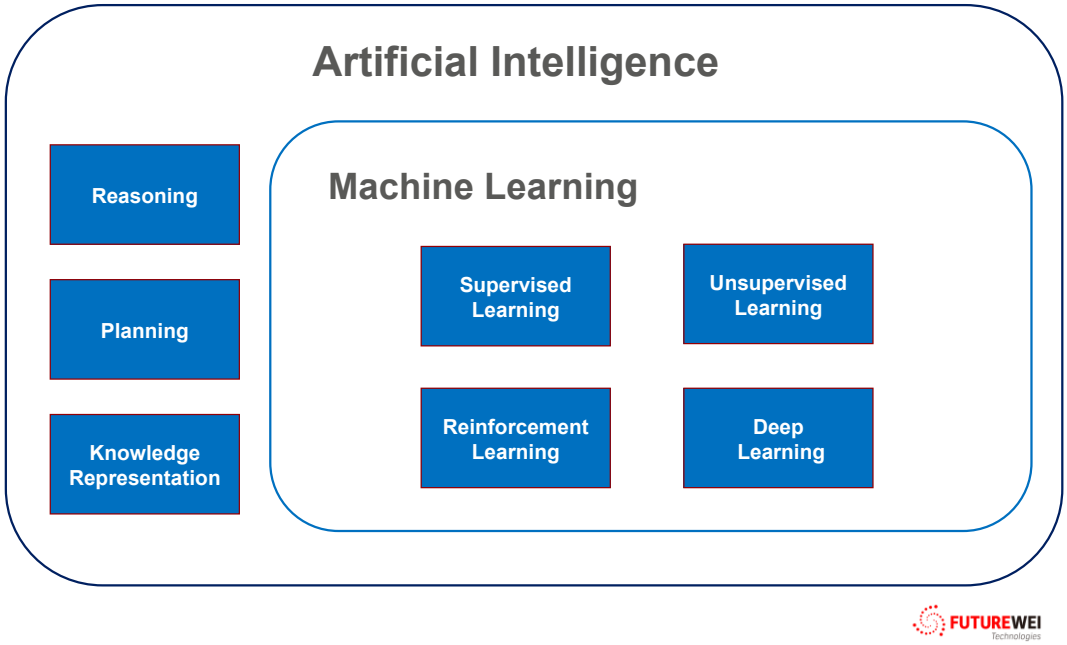
The AI ecosystem evolves around several key players. Nvidia, which has a dominant foothold on the latest GPU technology, is branching into software libraries and AI applications. Google, on the other hand, invested heavily in the open-sourced deep learning framework Tensorflow, while developing a proprietary Tensor Processing Unit (TPU) chip. Facebook is promoting PyTorch, which has obtained a similar market share as Tensorflow. Huawei developed Ascend, a Neural Processing Unit (NPU), and Mindspore, a deep learning framework.

Deep learning algorithms are evolving rapidly such that many capabilities of advanced AI programs 3 years ago are now considered basics. Thanks to the openness and sharing of the AI community, new concepts are adopted quickly into products. Many barriers to industrial adoption are no longer technical but legal and cost.

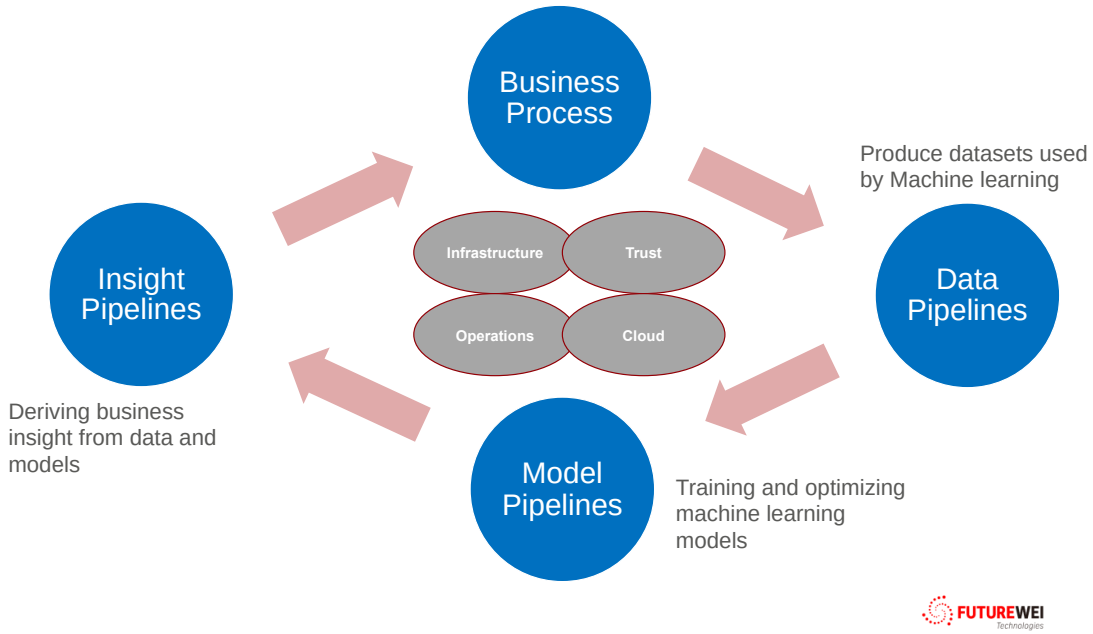
1.2 AI/ML WORKFLOWS, FRAMEWORK

AI/ML has gained extreme popularity over the last 10 years. However, Artificial intelligence research can be traced back to as early as the 1940s. Due to a large amount of dataset availability and extreme progress of hardware computation capability (especially GPUs), deep learning has gained tremendous traction. However, organizations are still leveraging traditional and other AI/ML techniques. It is still necessary to capture all the AI/ML technologies.

As shown in the following diagram, Machine learning is the key part of AI. AI, however, does cover other technologies such as reasoning, planning, and knowledge representation. DL has garnered more attention for the last 10 years due to the reasons mentioned before.



The differences between AI and ML are in semantics. ML is a technique and AI, on the other hand, is capability. AI means the capability of computers to demonstrate intelligent behavior. The realization of the AI capability involves many techniques and components, and ML is just one of them. Specifically, an AI system consists of three distinct building blocks: data pipelines, model pipelines, and insight pipelines. Data pipelines are processes that integrate siloed data sources and produce the datasets to be used by machine learning. Model pipelines are processing that train and tune machine learning models based on the datasets. Insight pipelines are processes that derive business insights from data analytics and model inferencing. Also, production AI must cope with infrastructure deployment and optimization. It must ensure AI decisions are trustworthy. It must operationalize the end-to-end system. And it must support hybrid multi-cloud environments.



All different components of an AI system shall be tacked to enable broad adoption of AI, not just the ML phase (Model pipelines).

1.3 INFRASTRUCTURE CHALLENGES

None of these innovations is possible without the help of an AI infrastructure. Like other IT infrastructures, an AI infrastructure consists of three major parts: compute, network, and storage. The compute part can be further divided into CPU-based and accelerator-based computations. The software stack can be divided into orchestration software, AI application software, and core AI system software.

The core AI software includes training infrastructure, inferencing infrastructure, framework (e.g., Tensorflow), distributed scheduling (e.g., Horovad [1]), and GPU-acceleration libraries (e.g., CUDA for Nvidia GPUs).

The application software includes applications that are used to solve real problems. For example, Nvidia Clara is an AI-powered healthcare application framework. There are many AI application software powering other industries, such as autonomous driving, life sciences, financial industry, etc.

The orchestration software is a component to facilitate the input/output of training, deploy the generated models, etc. This component is platform-dependent, and many orchestration software is built on top of infrastructure software provided by the platform. For example, if models are deployed as containers, then the AI orchestration software is using the container deployment infrastructure.

For cloud deployments, due to the size of the deployment, an AI infrastructure can be dedicated. However, for medium-sized corporations, a mid-size AI infrastructure may share hardware and software with big data infrastructure as both need access to storage and compute resources.

An interesting case for AI infrastructure is the hybrid cloud. Public clouds are known for their elasticity, which means a large set of compute or storage resources can be provisioned right away. The computing resources also include many software components that can be used to form a bigger application environment. Public clouds fit particularly well to use cases in which the workload fluctuates. Users can use dynamic provision and avoid

paying for any services that are not used. On the financial statement, users can save on the initial cost of building a data center and turn CAPEX into OPEX.

However, public clouds often face the cost issue when the scale is large and the workload is stable. In this case, buying or renting from on-prem IT vendors is often more cost-effective than using public cloud services. Another case of using on-prem services is when the user would like to use different hardware equipment from the ones provided by cloud vendors. In the AI infrastructure case, some users may prefer a particular Neural Processing Unit (NPU) or FPGA for acceleration, and this option is not available on the cloud.

Hybrid cloud has the potential to combine the pros of both public clouds and on-premise infrastructure. But the key issue to be solved is the data movement problem. It is well known that data gravity exists and customers are making different decisions based on their IT and financial needs.

Moving a large amount of data is very costly. Moving a large amount of data in a short time is even more costly. The network bandwidth is tight, and therefore the classic question was asked about whether it is faster to move the disks or transfer the data via a network. As a result, the center of the data gravity should be considered before the data started accumulation.

There are three types of hybrid clouds: (1) On-prem is the center and cloud is used as development to take advantage of the nimble of fast provisioning and development. But the production remains at the on-prem side for data safety and cost reduction. (2) The public cloud is the center and the on-prem side is used as a stage to make sure data is processed before it is transferred to the cloud. (3) On-prem is one of the multi-cloud centers. Workloads can be switched among these centers as needed.

In this document, we do not assume any of the three types. The computation (e.g., programs and models) can be moved from one data center to another at any time.

2 ENTERPRISE APPLICATION SECTOR

AI has been accepted into many industries and it is impossible to cover all the market leaders. Next, we referenced Matt Turck's "2020 data and AI landscape" diagram [2] and analyze applications among different sectors and industries. In this chapter, multiple sectors are analyzed. Several industries are surveyed in the next chapter.

2.1 SALES

2.1.1 Sector introduction

As big data and ML especially deep learning technologies becoming more and more mature and mainstream, customer relationship management becomes more data-driven. Using predictive analytics has grown into one of the major forces for sales. There are more and more sale-enabling software and service companies are leveraging those technologies to enhance customer relationship management, proactively help companies to reach new customers, and chose the right channel to effectively reach potential customers. CRM automation is greatly enhanced by leveraging AI/ML and big data. Sales teams can interact with customers and pursue sales opportunities not just relying on their sales mythologies on personal experience and anecdotal evidence but also using fact-based reasoning.

2.1.2 Key players

Key players include Salesforce, Chorus, Gong, InsideSales.com, AVISO, Conversica, Peole.ai, Clearbit, and tact.ai.

Salesforce uses deep learning, machine learning, and big data analytics platform (leveraging IBM Watson) to gain insights and provide predictive analytics for customer relationships. It is part of the Salesforce SasS platform offering. Chorus manages all unstructured data related to sales and uses ML to quickly spot which deals need attention to improve company win rates or sales cycle. It uses those data to facilitate and recommend customer communications. Chorus provides a "conversation cloud" that records, transcribes, and analyzes sales and customer success calls.

Gong company uses machine learning on data entry, gain a complete view of customer interactions, real-time data for decision making (Based on reality on opinions) and connect the entire company through a data-driven process.

2.1.3 Key Market Opportunities

Most of the companies listed here either deployed their own SasS cloud (like Salesforce) or leveraging public cloud vendors like AWS and Azure. It is cost-prohibitive for startup and small vendors to pursue home-grown AI technology and big data infrastructure, so they leverage public cloud vendors. Most of the companies surveyed here deploy their SaaS or services platform on public cloud platforms. Even they use on-prem training and deep learning platform to generate their home-brew AI algorithm, the market is too small for IT vendors.

However, providing big data and AI machine learning platform is still an opportunity for IT vendors to crack into this market sector as a supplement for the public clouds.

2.2 MARKETING

2.2.1 Sector introduction

Both B2B and B2C marketing have benefited from AI/ML and big data tremendously. As matter of fact, it is one of the earlier adopters for big data analytics and AI/ML. AI marketing use AI technologies to automate or at least assist market decisions based on data collection, data analysis. Those technologies can help customers to gain insight into their target audiences with more efficient channels. AI marketing, if used correctly, can significantly increase marketing ROI.

2.2.2 Key players

Key players include Salesforce, App Annie, D&B Lattice, 6Sense, Reflektion, Tubular, NGAGIO, TEALIUM, ACTIONIQ, Segment, Simon, Attentive, Active Campaign, SendGrid, ContenSqure, Zeta, Mparticle, Amerity, Bloomreach, Bluecore, INVOCA, and PERSADO.

2.2.3 AI applications for business

6Sense uses its AI platform to score over a quarter-billion accounts and people every day to reconstruct the account-based buyer journey for customer business, monitoring and analyzing changes in buyer intent at a massive scale. It leverages big data open-source software and tools to build its market platform. Those tools include Presto, Hadoop, Pymesos, Flume, etc.

2.2.4 Key Market Opportunities

CDP market alone is projected to be a multi-billion-dollar market. Most CDP vendors are leveraging similar big data, data warehouse, and data lake infrastructure. But it focuses on more identifiable individual data and leverage intelligence to provide more personalized content and delivery. Providing a unified data lake solution with intelligent content management, streaming process, data analytics platform, and ML platform may help market companies to create a more efficient and intelligent CDP platform. Since most public cloud companies are offering those solutions elastically, it is an uphill battle for on-prem vendors. Any vendors that can provide a hybrid solution will gain an advantage.

2.3 CUSTOMER EXPERIENCE

2.3.1 Sector introduction

Customer experience, or sometimes shortened as CX, is a key attribute to any customer-facing company. Many companies in this industry help companies not only gauge but also manage customer experiences (aka experience management or XM). As a big part of the world is digitized, the CX/XM industry is using more tools designed towards online market surveys, digital media, social media search, and AI/ML to process the information collected. Even though the market size is not as big as some other sectors, the customer experience management market is expected to reach USD 23.6 billion in 2027 and a CAGR of 17.7% [3].

2.3.2 Key players

Key players include IBM, Adobe, Qualtrics, Clarabridge, Surveymonkey, Zendesk, Pendo, Amplitude, and Kustomer.

2.3.3 AI applications for business

- **Customer feedback automation**
  The feedback process can be automated starting from when and where to send feedback requests. Then process the feedback automatically and generate insights.
- **Analytic tools**
  The analytics tools can use rule-based mechanism and ML mechanism to analyze unstructured data for customer experience-related contents and generate business information to support decisions.
- **Call center/Speech and natural language analysis**
  Analyze a large number of call records to extract “people-focused information” without human intervention.
- **Chatbots for customer service**
  Apply intelligent chatbots to support different customers.
- **Customer relationship**
  This part is overlapping with the marketing sector

Key players include IBM, Adobe, Qualtrics, Clarabridge, Surveymonkey, Zendesk, Pendo, Amplitude, and Kustomer.

2.3.4 Key Market Opportunities

This sector has large players, but many smaller companies are entering this market. A large part of AI is using traditional machine learning techniques, but new techniques such as NLP are the future of the industry. These products are software-based. Most products are provided as a service. For example, Clarabridge can be integrated into AWS Contact Center.

GDPR, HIPPA, and other regulatory requirements are a big factor in this industry because these services keep key corporate data. When building data centers, companies must meet the local data privacy regulations. Some big players, such as Qualtrics, built data centers across the globe. Others chose to build on public cloud vendors. For example, Zendesk is a partner of AWS and Pendo is a partner of Google Cloud.

It is an opportunity for IT vendors if on-prem or hybrid cloud solutions can be provided to these SaaS providers. As the industry grows, there might be a need for data center expansion and on-prem or hybrid cloud infrastructure. These infrastructures must have strong data privacy and RegTech capabilities.

2.4 HUMAN CAPITAL

2.4.1 Sector introduction

Human capital is one of the most important assets for companies to achieve goals, develop and remain competitive. Companies usually invest heavily in human capital through recruiting, education and training. The companies we focused on here are mostly focusing on providing a talent matching/searching platform that leverages AI and big data technologies. Most companies provide talent acquisition, talent management services through SaaS.

2.4.2 Key players

Key players include Pymetrics, AllyO, Mya, Wade, Texio, PI, HireVue, and FuseMachines.

2.4.3 AI applications for business

Pymetrics uses ML/NLP to analyze resumes, candidate data from publicly available websites, facilitating companies' talent search and clients' job searching process. Candidates will not be judged only by resumes but all other data available publicly. AllyO uses its AI recruiting platform to personalize candidate experience, automate the recruiting workflows and eventually give insights into your hiring process. A unique proposition is from FuseMachines – providing AI training for engineers, sales, and company personnel to retain the company's in-house talent.

2.4.4 Key Market Opportunities

The companies listed under this category are considered either small players or startup companies. It is cost-prohibitive for them to pursue home-grown AI technology and big data infrastructure, so they leverage public cloud vendors. Most of the companies surveyed here deploy their SaaS or services platform on public cloud platforms. Even they use on-prem training and deep learning platform to generate their home-brew AI algorithm, the market is too small for IT vendors.

However, providing big data and AI machine learning platform is still an opportunity for IT vendors to crack into this market sector as a supplement for the public clouds.

2.5 LEGAL

2.5.1 Sector introduction

At close to $1T globally, the legal services market is one of the largest in the world.

There are many similarities between the law and machine learning. For example, they both infer rules from historical examples to apply to new situations. Therefore the Law is conducive to AI and machine learning applications. The growing interest in applying AI in law is slowly transforming the profession and closing in on the work of paralegals, legal researchers, and litigators [4].

2.5.2 Key players

Key players include DocuSign, Kira, Ravel, Ross Intelligence, Lex Machina, and Disco.

2.5.3 AI applications for business

Machine Learning and NLP have enabled many AI tools to be developed to help legal departments to reduce costs, develop data-driven strategies, assess risk, and become more productive.

- **Contract Review and Negotiation**
  AI tools using NLP have been developed for the legal department to perform a legal textual analysis of proposed contract terms based on a legal department's objectives. AI tools serve as a check for a more effective review and identification of potential errors before contracts are finalized.
- **Contract Performance and Analytics**
  After parties have a contract in place, it is difficult to monitor contract performance to ensure terms and obligations are being met. NLP-powered AI tools allow legal extract key terms from the contract and compare those terms with a company's data metrics to determine whether terms and obligations are being met. Those tools help legal departments to harness the increasing data to assess contract performance and compile analytics.
- **Litigation Prediction and Analytics**
  AI tools help legal departments to predict the outcome of cases based on relevant precedent, facts of the case, and prior outcomes in particular jurisdictions. AI tools also predict the likelihood of success for motions or other pleadings based on data-driven assessments. These prediction models assist legal departments in making decisions on litigation strategies.
- **Legal Research**
  NLP-based AI tools can be used to build research platforms to uncover relevant law based on the fact pattern of a case. AI helps legal departments to review past matters to assess risks, potential liability and evaluate legal fee estimates based on analytics.

2.5.4 Key Market Opportunities

Although the legal profession is slow to adopt new methods. Law firms and in-house counsel have very little incentive to change as the current system has served them well over the years. But still, there are a lot of AI legal-tech companies are deploying technology in this sector, mainly to improve efficiency.

AI legal-tech companies usually use public cloud services (AWS, etc.). Bigger companies, such as DocuSign, have their own machine learning teams. Therefore, there is an opportunity for infrastructure vendors to provide AI and analytical systems to those companies.

2.6 RegTech & Compliance

2.6.1 Sector introduction

Regulatory technology (RegTech) uses information technology to enhance the regulatory process. RegTech is the management of regulatory processes. The main functions include regulatory monitoring, reporting, and compliance. The objective is to enhance transparency as well as consistency and to standardize regulatory processes. Compliance with regulations like GDPR, HIPAA, and the DFS Cyber Security Regulations is the largest subcategory of RegTech companies [5].

RegTech is often regarded as a subcategory under FinTech. With its main application in the financial sector, it is expanding into any regulated business with an appeal for the consumer goods industry.

The ongoing cost of regulation and compliance has become a primary concern industry-wide. On the expense side, post-crisis fines have exceeded $200 billion.

By 2027, global RegTech spending is expected to exceed $21.7 billion. According to a 2018 report by Medici, "an end-to-end implementation promises 634% in ROI realizable over three years".

2.6.2 Key players

Key players include IdentityMind Global (anti-fraud and risk management services), Trunomi (consent management to use customer personal data), Suade (submit requirement regulatory reports), PassFort (automates the collection and storage of customer due diligence data), and Compliance.ai (use ML to improve search, monitoring, and tracking of regulatory content).

2.6.3 AI applications for business

- **FCRM (Financial Crime Risk Management)**
  AI is applied in the areas of segmentation, data analysis, and scenario generation for efficiency and accuracy. AI is also used to augment decision-making, not replace it.
  RegTech tools monitor transactions that take place online in real-time to identify issues or irregularities in the digital payment sphere. Any outliner is relayed to the financial institution to analyze and determine whether fraudulent activity is taking place. Using big data and machine learning technology, RegTech reduces the risk to a company's compliance department on money laundering activities conducted online.
- **Financial Risk**
  ML and NLP are increasingly being applied to a broad range of problems within the areas of market risk, credit risk, and portfolio management. ML is used for scenario generation, curve construction, and validation. Together with NLP, ML tools are heavily used in data processing such as converting unstructured data to structured data. Clustering techniques such as topological data analysis and unsupervised neural networks are used widely to help setup factor analysis.
  Despite the wide adoption, AI tools are still seldom used in core models in financial risk.
- **GRC (Governance, risk, and compliance)**
  AI is used in automation and data validation, cleaning and speeding up the time-consuming elements that can stymie GRC systems. Another area is the use of NLP to help users with categorization and mapping.

2.6.4 Key Market Opportunities

AI applications in RegTech are in their infancy today. Many startups and AI vendors emerge in this area. RegTech companies collaborate with financial institutions and regulatory bodies, using cloud computing and big data to share information.

However, AI and big data platforms may be used by big financial institutions for financial risk management. Currently, their existing statistical model is still used for core applications. But they may go for the ML path any time soon.

2.7 FINANCE

2.7.1 Sector introduction

Artificial intelligence has given the financial industry a way to meet the demands of customers who want smarter, more convenient, safer ways to manage their money. AI is transforming the way we interact with money. AI is helping the financial industry to streamline and optimize processes ranging from credit decisions to quantitative trading and financial risk management [6] [7].

2.7.2 Key players

Key players include most commercial banks worldwide, infrastructure software platform vendor such as SAP S/4 HANA, application software platform and SaaS vendor such as Anaplan, and Zuora.

2.7.3 AI applications for business

Here are some key areas that AI/ML have been a proven help in the financial industry:

- **Fraud detection**
  AI has been very successful in battling financial fraud in the past decade or two — and the future is looking brighter every year, as machine learning is catching up with the criminals. AI is especially effective at preventing credit card fraud, which has been growing exponentially in recent years due to the increase of e-commerce and online transactions. Fraud detection systems analyze clients' behavior, location, and buying habits and trigger a security mechanism when something seems out of order and contradicts the established spending pattern.
- **Risk management**
  Artificial intelligence in finance is a powerful ally when it comes to analyzing real-time activities in any given market or environment; the accurate predictions and detailed forecasts it provides are based on multiple variables and vital to business planning.
- **Personalized Banking**
  Many financial apps offer personalized financial advice and help individuals achieve their financial goals. These intelligent systems track income, essential expenses, and spending habits and come up with an optimized plan and financial tips. Most big US banks are already offering such services to their customers.
- **Process Automation**
  AI-assisted robotic process automation for high-frequency repetitive tasks eliminates human error and cut expensive human workforce cost. Ernst & Young has reported a 50%-70% cost reduction for these kinds of tasks, and Forbes calls it a "Gateway Drug To Digital Transformation".

2.7.4 Key Market Opportunities

In recent years, there is a growing trend of moving applications to the cloud in the financial industry. While on-prem datacenter still plays a key role in the financial system, it is getting harder and harder. This trend is especially obvious for smaller players in the financial industry. The key reason behind the cloud transition is flexibility and cost (especially true for smaller companies).

Most big banks still run their on-prem data center; however, some are no longer expanding their on-prem capacity. There are limited opportunities in a few major banks that are still investing in the on-prem data center, these are the accounts that we should be focusing on. For them, initial cost, ease of use and maintenance, and security are going to be very important considerations. Besides large data centers, branch office needs may present another good market opportunity in the financial sector. HCI or dHCI may be well suited for such kind of environment.

2.8 AUTOMATION & RPA

2.8.1 Sector introduction

Robotic Process Automation (RPA) is the technology that allows people to configure the software to emulate and integrate the actions of a human interacting within digital systems to execute business operations. RPA robots interpret, trigger responses, and communicate with other systems to perform a vast variety of repetitive tasks. Essentially, any high-volume, business-rules-driven, repeatable process qualifies for automation.

RPA provides better accuracy, improved compliance, fast cost savings, super scalability, increased speed and productivity to many areas such as industry, banking & financial services, insurance, healthcare, manufacturing, hi-tech & telecoms, energy & utilities.

The enterprise RPA is growing at a CAGR of 65%, from 2016 to $3 billion in 2021. By 2021, Forrester estimates there will be 4 million robots doing office and administrative work as well as sales and related tasks.

2.8.2 Key player

Key players include UiPath, Blue Prism, Automation Anywhere, and Pega Systems.

2.8.3 AI applications for business

Intelligent Process Automation (IPA) is the application of AI and related new technologies (computer vision, cognitive automation, and machine learning) to RPA [8].

- **Computer Vision**
  Computer Vision is used in RPA tools to solve a variety of problems. Just as human beings do, they can see a screen or other scenes using contextual relationships. Computer vision can locate elements with configured search or coordinates.
- **Unattended Robotics**
  Autonomous automation, or robot-managing robots, is capable of monitoring and handling attended and unattended robot collaboration and optimizing end-to-end workflow automation with centralized work queues.
- **Cognitive Enhancements**
  Cognitive automation is an emerging field that augments RPA tools with AI such as optical character recognition (OCR) or natural language processing (NLP). With language detection, the extraction of unstructured data, and sentiment analysis, RPA can do knowledge-based processes. Language detection and sentiment analysis help the robots understand the meaning and emotion of text language and use it as the basis for complex decision-making.

2.8.4 Key Market Opportunities

Rn increasing number of PA vendors are using AI in the products they ship. Many RPA vendors integrated their end-to-end platform for hyper-automation with cloud infrastructure, cloud applications, and artificial intelligence solutions from a public cloud. For example, UiPath uses Amazon Textract, Amazon Rekognition, and Amazon Connect for dynamically-scaled AI-powered RPA solutions.

Many RPA vendors use third-party AI services from Google, IBM, Microsoft, and ABBYY. Those AI service providers could be potential AI and big data platform customers.

2.9 SECURITY

2.9.1 Sector introduction

Cyber-attacks are increasing in frequency, sophistication, and effectiveness. The ongoing trend of successful attacks demonstrates that legacy security systems do not keep pace with modern threats. This is because the traditional approach only detects well-defined threats. Under this new paradigm, AI technology is used to identify unseen cyber-threats at scale, in a variety of dynamic environments, in real-time, without human intervention.

The AI in the cybersecurity market is projected to generate a revenue of $101.8 billion in 2030, increasing from $8.6 billion in 2019, progressing at a 25.7 CAGR during the forecast period (2020 – 2030).

The market is categorized into threat intelligence, fraud detection/anti-fraud, security & vulnerability management, data loss prevention, identity & access management, intrusion detection/prevention system, antivirus/antimalware, unified threat management, and risk & compliance management, based on applications.

2.9.2 Key players

Key players include Trend Micro, Darktrace, blackberry, Norton, and IBM.

2.9.3 AI applications for business

- **AI Incident Prevention & Threat Detection**
  Supervised Machine Learning: Combining machine learning and old technologies (static and custom rules), the software can identify and block advanced threats. Unlike a rule-based model, an ML-trained model is aimed at rendering an entire class of attacks useless, essentially eliminating the need for hundreds or thousands of rules a security analyst would have to create and maintain to deliver comparable protection. It continuously hunts for threats without human intervention.
  Unsupervised Machine Learning: It can identify key patterns and trends in the data, without labeled-data training. Unsupervised machine learning is used to analyze network data at scale and make billions of calculations based on the current evidence instead of knowledge of past threats. It classifies data and detects compelling patterns. Based on this, it detects deviations from “normal” behaviors.
- **Automated Response**
  When a potential threat is identified, the software should take decisive actions in real-time to stop the attack and avoid the risk. AI and ML enable companies to reduce incident response time and comply with security best practices.
- **AI Analysis**
  AI Analysis can be used to triage incident reports from many customers. The supervised machine learning deploys experts’ knowledge on how analysts triage threatening and suspicious activities. It adapts to new and unique situations. AI analysis saves critical time and boosts productivity by allowing human experts to focus on strategic decision-making.

2.9.4 Key Market Opportunities

A new era in cybersecurity has begun. AI technology is the key trend in this area across diverse digital enterprises. All the security service vendors are investigating the AI approaches to detect, predict and analyze cyber-attacks. This presents a big opportunity for infrastructure vendors to provide AI platform solutions for them.

3 INDUSTRY APPLICATION SECTOR

3.1 ADVERTISING

3.1.1 Sector introduction

The advertising industry is a long-time embracer for AI. The more accurate the advertisement is, the higher price the buyers are willing to pay. As more people are connected to the Internet, the battlefield switched from traditional paper media to online media. With the help of two-way communication, the advertising industry can collect more data about users, such as feedback, mouse or finger movement, or even facial expressions. The pay-per-click (PPC) scheme gradually became mainstream for online advertising. Since businesses are not paying anything until the viewer clicks on the banner, advertising platforms need to target the viewers accurately.

Social media companies and search engines have the advantage to see what people are interested in so that they provide context-aware advertisements. The advertising market is fueled by technologies and many high-tech companies are benefiting from it. For example, Google’s advertising division revenue was USD 37.1 billion in 2019 [9]. Surrounding the big companies, many players are trying to match the advertisers and the publishers.

There are several main roles in this market:

- **DSP, or demand-side platform**
  DSP represents the advertiser (aka “buy-side”). An advertiser can use a DSP to access multiple ad exchanges, track the performance of the ads, and optimize the performance. The value of DSPs is to help advertisers to reach certain target audiences with minimal buying cost. For example, Google Ads can work as a DSP to give advertisers to buy “impressions” from Google’s search results and other suppliers.
- **SSP, or supply-side platform**
  SSP is a service that represents the publisher (aka supplier, partners, or “sell-side”). An example of a publisher is a website hosting sports news. A publisher can use an SSP to place online ads, manage them, and optimize them. The value of SSPs is to maximize the selling prices. For example, Google Adsense can work as an SSP.
- **Ad exchange and ad network**
  An ad exchange is where the suppliers can list their inventories (aka slots) and transactions with advertisers. An ad exchange can interact with SSPs or with suppliers directly. A strict role in ad exchange is where bidding happens.

Note that many companies do business with many sides and thus loosely speaking an ad exchange company may serve as both DSP and SSP. DoubleClick is such an example.

3.1.2 Key players

Key players include Google, Facebook, DV360, Amazon, DSPs and SSPs, Xandr, MediaMath, Criteo, Integral Ad Science, Albert, Gumgum, Appier, The Trade Desk, Marketing SaaS Cloud vendor, Oracle Data Cloud (formerly Moat), and Salesforce Datorama.

3.1.3 AI applications for business

- **Advertisement platforms (including social media platforms)**
  On the buyer side, DSPs are using AI to predict the prices and bidding success rates. The platform looks at millions of queries per second and plans the next action.
- **Ad targeting**
  On the seller side, SSPs are collecting user behaviors based on browser cookies, website visited, and search keywords, etc. Then users are classified into different clusters in terms of user interests. For example, SSPs will display car ads to users who are recently browsing car dealer websites.
- **Budgeting and insights**
  DSPs can use ML to help advertisers to predict budget spending and make sure the spending is on track and the return is as expected.
- **Content creation and personalization**
  The ad's content can be dynamically created based on the results of the target group and merchandise. The layout, color, wording, and many other features could be customized and individualized to improve the click rate.

3.1.4 Key Market Opportunities

The need for infrastructure is quite big in the advertising industry because the queries can be millions per second. For accurate results, the back-end side must generate the latest strategy to cope with market changes. Many key players are international. Depending on the stage of the company, it may go to a public cloud vendor to satisfy the infrastructure needs or build on-prem data centers. The accurate ratio is unclear, but it appears that vendors such as AWS takes a large chunk of the market. Many companies surveyed in this document are

AWS partners. Besides IaaS vendors, they may use SaaS providers. For example, Gumgum is using Databricks and delta lake for its analytical platform. The need for storage is big as Gumgum is processing 50TB of data per day. Streaming data processing and its performance is critical. But batch processing is also important.

New market opportunities include providing solutions to customers so that they can build an cost-effective marketing platform using on-prem, public clouds, and hybrid clouds. The on-prem solutions should focus on mid to large market SaaS providers.

3.2 EDUCATION

3.2.1 Sector introduction

The education sector has been stable for a long time as the student size of universities and K-12 schools are quite stable. However, the online learning firms and training firms started to the public needs for continuous education. Other than Massive Open Online Courses (MOOCs), many universities also joined the trend to provide online learning. The size of the smart education and learning market may surprise many people. According to Grand view research [10], the smart education and learning market will worth USD 680 billion in 2027. In 2019 [11], the market size is USD 182 billion, and the compound annual growth rate (CAGR) of 17.9%.

Although the boldest prediction is that AI might replace teachers one day, most scholars believe that teachers are a necessary role. Like many other industry sectors, AI might become assistants to humans, in this case as teachers and administrators.

3.2.2 Key players

Key players include Nuance, Grammarly, Knewton, Cognii, Querium, Century Tech, Carnegie Learning, Squirrel AI learning, and Liulishuo.

3.2.3 AI applications for business

- **Individualized learning**
  Machine learning and recent progress in cognitive science together can plan individualized learning for each student. Feedbacks can be obtained from students' testing, speed of progression, and even their facial expressions. AI can assist students to gather extra instructions and materials, like human tutors.
- **Language learning/critique**
  With the latest progress in Natural Language Processing (NLP), AI can help students learn a new language in ways of listening, speaking, reading, and writing. A bot can in large percent replace a language tutor and human educators can work on more educational tasks.
- **Smart administration**
  Administration tasks, such as scheduling and optimization, can be offloaded to AI. So that administrators can focus on education and policy issues.

3.2.4 Key Market Opportunities

Like in many other sectors, the new players in smart education often adopt public clouds as their infrastructure. For example, Knewton is using AWS EMR for its big data infrastructure.

A larger size company may opt to use on-prem data center solutions. Particularly companies that are hosting their cloud services. However, AI usage in education is still in its beginning stage. The AI training infrastructure is still limited to basic machine learning and big data. The vision of having a virtual classroom is valid but still under construction. Even though at this stage, other than common big data infrastructure no special AI infrastructure is needed, IT vendors can partner with education providers and explore new applications with them.

3.3 REAL ESTATE

3.3.1 Sector introduction

According to Morgan Stanley Digitization Index, real estate is the second least digitized industry in the world. The real estate sector has always been slow to adopt new technology. However, in recent years, many companies started to recognize the immense potential of AI. AI technology is now applied to real estate to improve clients' home search, identify strong lead gen, remove bias from recruiting, refine the transaction and better predict market values [12].

3.3.2 Key players

Key players include Redfin, Zillow, and Rex.

3.3.3 AI applications for business

- **Home Search**
  AI-powered algorithms identify the user's preferences and suggest property based on its findings. For example, Trulia's app uses computer vision to extract relevant information from the user's photos and shows best-matching offerings on top of the search results, and recommends other listings accordingly. It also considers the preferences of other users that looked at similar properties which allow for a superior level of personalization.
- **Predict Property Value**
  AI tools can anticipate rent and sale price fluctuations or identify the perfect timing for selling a property. Skyline (Israeli startup) AI uses predictive analysis to accurately assess property value. Zillow found another use case to partially estimate property value by analyzing photos. Machine learning techniques can assess even the most sophisticated interior details that sell it to the customer.
- **Real Estate Management**
  For large properties like corporate office buildings, the maintenance cost can take up a significant part of the total budget. IBM AI-powered TRIRIGA solution helps management professionals to effectively utilize office space by gathering data from sensors and analyzing it with an AI algorithm. Employees can talk to spaces with the help of natural language processing and AI tools can identify users' needs and rearrange the entire office layout.
  AI technology can also be used in energy-saving and property resource optimization. Gridium's ML algorithm automatically analyzes weather data and detects energy use patterns to warn property manages.
- **Lead Generation**
  AI can accurately identify potential buyers by their activity from those who are browsing out of curiosity. Moreover, the algorithm can also identify what type of property the customer is looking for. This helps agents to save time and effort in dealing with customers.
- **Mortgage Lending**
  Mortgage lending is data-intensive with bank statements, credit history, proof of income, and many other papers. The mortgage lending sector currently uses optical character recognition (OCR) to help lenders automatically read data from borrowers' documents. Unfortunately, most of the documents are unstructured and need human validation. Machine learning tools can capture significantly more information with higher accuracy and less human interference. Combining OCR and ML tools, borrowers can process more requests with a better experience.

3.3.4 Key Market Opportunities

AI is rapidly penetrating real estate software development. The challenge is that the majority of the accumulated data remains siloed and lacks standardization. These data need to be made interoperable before AI technology can be reaped in this sector.

For real estate software vendors, those who deploy various metrics using AI technology will stay ahead of the competition. It is a big opportunity for infrastructure vendors to provide big data and AI platforms to those vendors.

3.4 GOVERNMENT & INTELLIGENCE

3.4.1 Sector introduction

Applications of AI to the public sector are broad and growing, with early experiments take place around the world. The use of AI in government comes with significant benefits, including efficiencies resulting in cost savings, and reducing the opportunities for corruption. For example, Deloitte has estimated that automation could save US government employees between 96.7 million to 1.2 billion hours a year, resulting in potential savings of between $3.3 billion to $41.1 billion a year [13].

3.4.2 Key players

Key players include Palantir, Opengov, Dataminr, and Anduril.

3.4.3 AI applications for business

The potential uses of AI in government are wide and varied. Mehr suggests that six types of government problems are appropriate for AI applications:

- Resource allocation: such as where administrative support is required to complete tasks more quickly.
- Large datasets: where these are too large for employees to work efficiently and multiple datasets could be combined to provide greater insights.
- Expert shortage: where basic questions could be answered and niche issues can be learned.
- Predictable scenario: historical data makes the situation predictable.
- Procedural: repetitive tasks where inputs or outputs have a binary answer.
- Diverse data: where data takes a variety of forms (such as visual and linguistic) and needs to be summarized regularly.

Potential and actual uses of AI in government can be divided into three broad categories:

1. **Contribute to public objectives**
   There is a range of examples of where AI can contribute to public policy objectives:
   - Receiving benefits at job loss, retirement, bereavement, and childbirth almost immediately, in an automated way (thus without requiring any actions from citizens at all)
   - Social insurance service provision
   - Classifying emergency calls based on their urgency
   - Detecting and preventing the spread of diseases
2. **Assist public interactions with government**
   - Answering questions using virtual assistants or chatbots
   - Directing requests to the appropriate area within the government
   - Filling out forms
   - Assisting with searching documents
   - Scheduling appointments

3. **Other uses**
    - Translation
    - Drafting documents

3.4.4 Key Market Opportunities

While applications of AI in government work have not kept pace with the rapid expansion of AI in the private sector, the potential use cases in the public sector mirror common applications in the private sector. However, potential risks associated with the use of AI in government include AI is susceptible to bias, a lack of transparency in how an application may make decisions, and the accountability for any such decision.

AI application vendors for the government should focus more on security, transparency, and accountability. AI infrastructure vendors should also meet similar requirements for the AI and big data platforms.

3.5 COMMERCE

3.5.1 Sector introduction

As machine learning especially deep learning gaining momentum in recent years, more and more businesses are looking for ways to take advantage of what AI can bring to their business. There are a few trends in retail and commerce areas. Big retailers such as Amazon, Walmart, and Costco have dominant positions in this sector, which leave other players to scramble and leverage even more on AI to grab some market shares as niche players.

Stich Fix leverages AI and data science to actively push products individual clients may like. This customized and personal shop consulting model has gained some loyalties of their clients.

HowGood has been collecting and managing a large amount of environmental data and using AI and data science to answer their client's questions about the environmental and social impact of their products.

Faire use AI and Machine learning to predict spending patterns and analyze customer cash flows, credit score, etc. It provides such quick credit approval services to other retail, financial, and commerce companies.

3.5.2 Key players

Key players include FAIRE , STITCH FIX, and HowGood.

3.5.3 AI applications for business

By answering social impact and environmental impacts, AI and data science can help companies customize the personal shopping experience. Big data and AI provide instant credit scores and credit approval.

3.5.4 Key Market Opportunities

The companies listed under this category are considered either niche players or startup companies. It is expensive for them to build home-grown AI and big data infrastructure without leveraging public cloud vendors. Most of the companies deploy their SasS or services platform on public cloud platforms. Even they use on-prem training and deep learning platform to generate their home-brew AI algorithm, the market is too small for IT vendors.

However, big vendors like Walmart, STIME, Carrefour may present better opportunities for IT vendors. But it is out of the scope for this white paper.

3.6 FINANCE – LENDING

3.6.1 Sector introduction

AI has been widely adopted in many corners of the financing industry, including one of the most important sectors – the lending business. AI can help from multiple angles in the lending industry, including but not limited to:

- Risk assessment
- Fairness of loan
- Finding potential clients
- Deciding loan size and rates

The traditional way of loan assessment involved many parameters that were pre-built by the banking industry based on past experiences, which are frequently biased or outdated. It takes a tremendous amount of data and effort to change the existing loan model. AI's top strength is building a model based on the huge amount of data, and even better – dynamically update the model based on new incoming data. This is exactly what is needed by the finance industry, especially so for the lending sector.

3.6.2 Key players

Key players include most big commercial banks such as Bank of America (they have been shifting on and off the cloud, but always kept core compliance data on-prem), Zest finance, OnDeck, Affirm, and Ppdai.com.

3.6.3 AI applications for business

The existing bank lending assessment model is based on a slew of parameters including but not limited to – income, race, asset size, loan term, location, business sector, etc. This model was built over the past tens of years and has been treated like the bible of the finance industry. It is not easy to change them, however, they are surely outdated and frequently biased already. Many small banks and startups have already started practicing AI-assisted lending business, while big banks are typically more conservative and mostly still on the watch.

The most important application of AI in the lending business is to build a dynamic model for data-driven decision-making for loans, the model can help evaluate credit risk, and decide the amount of loan and a fair interest rate to match the risk.

3.6.4 Key Market Opportunities

Due to the extremely high security, privacy, and compliance requirements, banks are typically not going to use cloud service for their core operations, including the loan assessment platform. During the CapitalOne security breach [14] in 2019, the financial industry questioned the validity of the "cloud-first" strategy. The latest trend appears to be favoring a hybrid cloud model, with on-prem data centers hosting core data and services, while the public cloud offers flexibility and on-demand expansion. We believe there are tremendous business opportunities in this sector for selling on-prem equipment that can facilitate their AI application platform.

From a storage business point of view, the most suitable product for this sector is a high-performance, AI-capable storage platform, with plenty of security and privacy compliance features and capabilities embedded. A mid-range storage system with SED encryption, combined with AI capable compute nodes should fit the majority of potential customer needs in this sector. All involved components must meet popular security and privacy regulations such as GDPR.

3.7 FINANCE INVESTING

3.7.1 Sector introduction

Wall Street has always been the earliest adopter of any latest AI/ML progress since any breakthrough or innovation in this field could easily translate into billions of profits on Wall Street. Hedge funds, investment banks, and many quantitative analysis companies are always looking for the fastest and most advanced AI/ML algorithms and applications, as well as the best AI platform that can host them. There are tremendous AI and storage platform business opportunities in this industry.

3.7.2 Key players

Key players include many hedge funds on Wall Street, major investment banks such as Goldman Sachs, Morgan Stanley, and some startups offering AI quant services, such as Quantopian, and Cognizant.

3.7.3 AI applications for business

Quantitative analysis is the most important tool on Wall Street in recent decades. It has largely replaced the role of human traders in many institutes. The two most important way that ML/AI can help in investing industry are:

- Scan and watch for trading patterns of stocks and other investment items to dig out good candidates. This is like OLAP, but with extensive AI/ML for recognizing complex patterns.
- The huge benefit of removing human emotion from decision-making has been proven to be extremely helpful in predicting market directions and making trading decisions. This is somewhat like OLTP with extensive trading signal processing and pattern recognition with the help of AI/ML. The response of AI applications is much faster and precise than human observation when things happen.

3.7.4 Key Market Opportunities

Hedge funds are frequently trading at high frequency, sometimes even milliseconds matter, so the cloud is normally not their preference. They always demand the best hardware platform to run their applications on-prem. For this market, high-end storage and AI compute platform is recommended. The proposed solution should include all-flash storage with the lowest possible latency and high throughput, with Nvidia DGX like AI/ML compute nodes. RDMA fabric between the compute and storage nodes can offer the best performance when storage and compute nodes need to communicate.

3.8 INSURANCE

3.8.1 Sector introduction

AI can potentially change the insurance industry dramatically, but so far it has only scratched the surface. There are dozens of processes that could be greatly improved using AI, and over time, more insurance providers adopt AI in more areas. Some of the applications that can benefit most from AI include pricing, claiming, and fraud detection [15] [16].

3.8.2 Key players

Key players include major insurance companies, EY insurance, Metromile, Lemonade, Zesty.ai, Cyence, and Hippo.

3.8.3 AI applications for business

Here are some examples of many areas where AI can help the insurance industry:

- Behavioral Policy Pricing
- Risk Assessment Model

- Customer Experience & Coverage Personalization
- Faster, Customized Claims Settlement

These can be summarized into a few categories as mentioned in the introduction: pricing, claims-handling, and fraud detection:

- **Pricing**
  One of the most promising ways AI can improve the insurance industry is around pricing. With AI, insurance companies can easily offer personalized pricing based on each customer's personalized risk assessment. With the personalized model, insurance companies can provide competitive prices to "good" customers and avoid high-risk customers that are not even worth the "market price".

- **Claim handling**
  Insurance companies spend a lot of money on claims personnel, and insurance prices are often marked up to account for case-solving. A large portion of these processes around claims management and payouts can be automated by AI. The saved hiring cost can be a good chunk of the whole insurance industry revenue.

- **Fraud detection**
  Insurance fraud costs the industry more than $40 billion per year. AI is an effective way to detect fraud and prevent risk. Using AI can help insurance companies spot abnormalities in claims data and identify false information that customers use to get a lower premium or bigger claim payout.

3.8.4 Key Market Opportunities

The insurance industry is a huge business sector. Even though AI penetration is not very high in this sector yet, we do see very promising fields for both large and small insurance companies. Within large companies, there will also be demands for both central datacenter and branch offices. Storage and compute solutions for the insurance industry need to cover both ends of the size spectrum – both large datacenter and small edge nodes. On the compliance side, since the insurance industry deals with lots of personal data, it has very strict security and privacy requirements like the finance industry. All components in our proposed solution need to meet the most popular privacy standards such as GDPR. For large companies, A scalable solution with multi-site redundancy should satisfy their data center storage needs. For edge office and small insurance companies,

3.9 HEALTHCARE

3.9.1 Sector introduction

Healthcare is one of the most important sectors in North America and the European economy. Healthcare spending is around $4 trillion per year in the U.S. Healthcare expenditure amounted to around more than 7 - 10% GDP in the EU and around EUR 944B [17].

However, like most bureaucratic huge enterprises, Healthcare costs kept rising with no end. One of the major reasons is the compound factor of rising health professional cost, administrative cost, etc.

AI/Machine learning and healthcare are in many aspects are a good combination for one another. Healthcare in its core is pattern recognition. Machine learning excels at pattern recognition and processing a large amount of data. Given enough data sets, it can "learn" enough knowledge to handle a lot of tricky works previously only being handled by healthcare professionals.

Healthcare industry can be broken down into 3 categories: clinical (the delivery of care to patients), administrative (the operational nuts and bolts that keep the healthcare system running), and pharma (the research and development of new medical drugs).

We will breakdown those categories in the "AI applications for business" section.

3.9.2 Key players

Key players are divided into 3 major categories below:

- Clinical
    - Caption Health, PathAI, Paige, and Zebra Medical Vision
    - Babylon Health
    - Biosourmis, Current Health, Myia, Aluna
    - Syapse, GNS Healthcare, Tempus
    - Flatiron
    - Metabiota(Epidemic tracking platform)
    - Linkdoc
    - Imagen(Medical image analysis)

- Administrative
    - Olive, Notable Health,
    - Alpha Health(RCM)
    - Qventus(Patient flow)
    - Protenus(Regulatory Compliance)
    - Kmodo Health, Datavant, Abacus Insights, healthVerity, Kyruus, Ribbon Health, Redox(Data Infrastructure)
    - SukiEpic, Center(Medical Documentation)
    - Kyruus(Provider access, Patient Access)
    - Zebra(Mobile Clinical)

- Pharma
    - Novartis, Verge Genomics (New drug discovery)
    - Bayer, Merck
    - IBM Watson, GNS Healthcare
    - Deep 6 and Antidote (Data management)

3.9.3 AI applications for business

- **Clinical**
  Using computer vision to identify health conditions in medical images has become the most widely referenced use case for AI in healthcare. There are the following major categories in this area:

  Imaging, Patient Intake and Engagement, Remote Health, In-Hospital Care, and Precision Medicine.

- **Administrative**
  Compared to clinical or life sciences use cases, applying AI to the administrative side of healthcare may seem unglamorous. But an enormous opportunity for value creation exists here. The Healthcare system is plagued by waste and inefficiency. AI can be leveraged in those areas including Provider Operations: RCM, Patients flow and resource allocation, regulatory compliance (GDPR, HIPAA, etc.), data Infrastructure, and Medical Documentation.

- **Pharma**
  AI can significantly help the Pharma industry to perform tasks that traditionally rely on human intelligence. Over the last 5 years, the use of AI in the pharma and biotech industry has redefined how scientists develop new drugs, tackle diseases, and more.

  Computer vision can reduce drug discovery time to screen and predict which untested compounds might be worth exploring in more detail. NLP can be used to screen patients for drug trial enrollment.

  Most pharma companies' current IT infrastructure is based on legacy systems that were not designed with AI in mind. They lack sufficient data storage and often lack interoperability. Most data within medical systems are in free form, until there are data management software made them readily available, the information cannot be processed and used efficiently by health professionals.

3.9.4 Key Market Opportunities

Healthcare is moving to a new outcome-based model. Data is going to be increasingly important and more valuable. How to create, manage the multitude of source data, and normalizes data from data sources is very critical to the success of this model. Healthcare companies are increasingly using AI in their analytics engine to pull relevant insights from data such as EMR to generate more value to clinical, healthcare administrative, and drug discovery, etc.

Providing a big data platform and enabling managing massive data for AI training are still the key requirements to this sector. That certainly also includes regulation conforming (GDPR, HIPAA), data encryption, etc.

Due to the nature of Healthcare companies (a lot of new startups in this area to leverage AI capability), most smaller players are creating their clouds based on public cloud vendors like AWS, Azure.

Flatiron health uses AWS for its analytical applications and data processing platform OncoCloudTM software to AWS. AWS offers some major features including data encryption, identify management, strict compliance, and additional auditing and security capabilities besides AI and computing capabilities.

Although on-prem and private cloud vendors may provide similar technology and scale to meet the requirements, healthcare companies (especially small to medium size) are increasingly looking for public cloud vendors. The nature of public clouds gives those companies fast starts with minimal up-front costs.

However, due to regulation differences among countries, some healthcare companies are providing traditional software license-based products (Linkdoc). The opportunities for storage and computing are resting on providing open source-based big data and machine learning solutions.

Bigger pharma has accumulated a large amount of data over the years and has been looking at ways to explore and unlock those data values by leveraging AI technologies. It also presents big opportunities for infrastructure vendors to provide solutions for them to deploy AI platforms on top of big data lake infrastructure.

3.10 LIFE SCIENCES

3.10.1 Sector introduction

The rise of AI changed the landscape of many industries and research fields, life science is one of the pioneers in adopting AI. Although we are still in a very early stage of applying AI in life science, there are already tremendous benefits observed in some areas. There are pockets of early adopters trailblazing new approaches

and seeking a competitive edge to accelerate products to market, improve patient outcomes and care, and drive cost efficiency [18].

In recent years, life science and medical research advance made important breakthroughs in some key areas, such as genome sequencing, gene editing, personalized medicine, AI-assisted drug screening, etc. With the application of these new technologies, medical data is growing at an explosive rate. For example, a single patient's genome sequencing data can require tens of TBs and need to be retained for many years. To deal with these huge data and analyze them for clinical use or research purpose, AI and big data technology are crucial. A report by Accenture estimates that by the year 2026 this market will grow to $150 billion/year. The main purpose of AI or machine learning applications specific to life science is to make data accessible and usable for improving prevention, diagnosis, treatment, and research.

3.10.2 Key players

Key players include Merk, Pfizer, 23andme, iCarbonX, WiXiNextCode, Verily, Zymergen, and Pathway Genomics.

3.10.3 Applications for business

Here are some key fields where life science can greatly benefit from AI technology:

- **Advancing Diagnostics**
  Image recognition is where AI started gaining momentum in recent years, and this happens to be one key area for disease diagnose. Histopathology image analysis and automated diagnosis were some of the earliest fields that benefited from AI technology. AI and pattern recognition, combined with complex algorithms and automated immunohistochemical measurement systems, have advanced pathologists' ability to oversee the analysis and concentrate on more difficult cases. In some cases, AI can already achieve better precision than an experienced doctor in reading medical images for diagnosis. Although the human review is still required due to regulation, AI is providing a fast and reliable reference for a more efficient diagnosis, thereby greatly reducing the cost.

  In addition to replacing doctors to interpret the medical image for disease diagnose, AI has also enabled some previously impossible applications, such as portable MRI. Traditional MRI requires good quality images in huge quantity to draw a clear conclusion, so portable MRI devices could not satisfy the needs due to their relatively weaker magnetic field and stability. However, with AI's strong ability to find patterns among the noise, portable MRI is becoming feasible.

- **Advancing Research of New Products**
  A key strength of AI is digging useful patterns from a large amount of data, more and more companies are taking advantage of this data mining capability to help identify the new potential of existing products or identify new candidates. Examples include, but are not limited to:
  - Uncover insights that can lead to the identification of new mechanisms of disease, potential new line extension, and design for preclinical experiments.
  - Understand the interaction between small-molecule medicine and the target protein.
  - Extract relevant knowledge from commercial, scientific, and regulatory literature, allowing researchers to identify competitive white space, eliminate blind spots in research, and discover disease similarities.

- **Accelerating Drug Development**
  Drug development is an extremely long and costly process, frequently taking close to 10 years, with sights set on reducing them to five to seven years. Advancements in AI and machine learning can help in multiple stages of this long process, starting from drug candidate screening. Scientists are integrating research data, lab data, and clinical data, etc. to create a holistic picture of the drug development candidate and use AI to help mine the data in real-time to help make improved decisions faster, which will accelerate the product development and scale-up process.

- **NLP Driving Compliance in Clinical Trial Compliance**
  The clinical trial is arguably the most lengthy and costly part of the drug development process, frequently even more demanding than the drug development and lab test phase. Compliance is a big part of the problem. New applications based on NLP technologies incorporating scientific-specific taxonomies and text-mining models are emerging. It is possible to identify keywords, phrases, and data patterns that may require redaction or anonymization. These new applications provide the higher level of accuracy required to meet the policy requirements while also automating manual activities.

- **AI Model Improving Clinical Trial Operation**
  Nearly 80 percent of clinical trials fail to meet their patient enrollment deadlines. AI models can improve and accelerate clinical site and patient selection decisions by highlighting high-probability targets. It can benefit the clinical trial in both quality and speed, hence greatly lower the cost and increase the success rate.

3.10.4 Key Market Opportunities

It is worth noting that there are many small startups actively engaged in life science research and new drug development, as well as many academic researchers. They typically have a very small footprint but need to deal with a large amount of data since many baseline data sets are the same for a wide range of research, such as existing drug databases, human genome databases, etc.

For large corporations like Merk and Pfizer, the biggest issue they are facing is data silo within the company and a large quantity of duplicates and inefficiency as a result of the isolation. For these businesses, a solid data lake solution with an integrated data analytics platform can be very valuable and may help them reduce cost and improve efficiency.

For small companies, startups, and research organizations, the cloud is frequently their 1st choice to get started when there are many uncertainties. However, once they reach a steady operational state, the high cost of using the cloud could push them to consider an on-prem solution. The ability to seamlessly integrate with the cloud and the agility of the infrastructure will be crucial for them when picking the on-prem infrastructure. A scalable and highly integrated compute plus storage with a preloaded AI software stack would be ideal for them. We should consider HCI based solution with GPU/NPU capability and a cloud connector.

3.11 TRANSPORTATION

3.11.1 Sector introduction

According to the Grand View Research report [19], two years ago (2019) the global intelligent transportation market size was USD 26.58 billion, and the estimated compound annual growth rate is 5.8% from 2020 to 2027. This is a growing market and intelligent transportation is only a small percentage of the global transportation market, which is about USD 5 trillion [20].

The transportation industry is embracing AI to transform from human driving to driver-assisted driving, and eventually to autonomous driving. There is a broad range of companies pushing for innovations for our daily transportation. The list includes automobile manufacturers, ride-sharing companies, and autonomous driving development companies. Some companies focus on personal trips, while some others focus on cargos or mining. The development of AI has the potential to replace human operators with computers and bring revolution to the whole industry.

The ride-sharing companies are trying to use AI algorithms to change the way people move. First, they use AI algorithms to optimize the matching between ride providers and passengers. Then they are developing self-driving cars and buses to provide 24x7 riding services and lower the cost of each riding trip.

Car manufacturers, on the other hand, are investigating self-driving cars for a different reason. Passengers would like to be freed from driving as labor and spend the time on the vehicle for leisure, such as reading, watching shows, or taking a nap. Some car owners would like to minimize the attention needed for their cars and use them only as transportation tools. Instead of 100% own a car, some car owners may want to switch to the ride rental model, and thus converge with the ride-sharing companies' vision.

Automobile companies are also investigating self-driving vehicles that are not for humans. For example, self-driving trucks are ideal for shipping goods from point A to point B. The mining industry is interested in self-driving mining trucks that can maneuver in hilly terrains and ship mines back and forth.

The recent trend of electric vehicles makes autonomous driving more realistic because electric vehicles have less complicated mechanical components than traditional vehicles. Many of the components can be controlled by software. Thus, some industrial experts predict that vehicles will become more like our smartphones and we should expect software updates for more features and bug fixes.

Therefore, as the need for self-driving vehicles grows, many autonomous driving development companies are aiming to provide solutions that can serve multiple automobile manufacturers and ride-sharing companies. They work with automobile manufacturers and provide the tools and platforms needed.

3.11.2 Key players

Key players are divided into following categories:

- Vehicle manufacturers (passenger vehicles and mini-buses): Tesla, Volkswagen, Toyota, NIO, Cruise/GM, Zoox, Optimus, NURO
- Ride-sharing companies: Uber, Lyft
- Autonomous driving development companies: Waymo, Argo (Ford, VW), AImotive, Aptiv, Perceptive Automata, Aurora, Kodiak, Tusimple

3.11.3 AI applications for business

- **Autonomous driving and Advanced Driver Assistance System (ADAS)**
  The challenges for ADAS are obvious. The safety of passengers, other drivers, and pedestrians on the road must be guaranteed. The reliability of autonomous driving must be proved under different terrain and weather conditions. These requirements brought challenges to the design of an AI infrastructure for Advanced Driver Assistance System (ADAS) software development.

- **Ridesharing applications**
  Ridesharing companies are using AI to improve ride-sharing applications, both on the user side and the backend. On the user side, AI can give customers a good riding experience with more accurate estimated time to arrival (ETA) prediction, hands-free communication, passenger safety, and a low

waiting time. On the backend, AI algorithms can help the companies improve scheduling efficiency by better predicting rider demand and ETAs.

Many AI techniques can find their usage in a ride-sharing application. Traditional operation optimization can be used for optimizing efficiency. Neural networks can be used to predict vehicle ETAs. Natural language processing can be used for customer interfacing. Computer vision is used to authenticate the driver's identification. We expect the usage of AI to expand into future ride-sharing applications as well.

Ridesharing companies and their applications are also contributing to the AI community via new algorithms and open-source infrastructures. Due to the large amount of data collected, new algorithms are invented and tested in a real-life testbed. They are also sharing their infrastructures. For example, Uber open-sourced Ludwig [21] -- a code-free AI toolbox built on top of Tensorflow, Neuropod [22] -- a middle layer to use multiple AI frameworks, and Horovad [1] [23] -- a distributed deep learning training framework. Lyft open-sourced Flyte [24], which is a distributed platform for constructing machine learning workflows.

As the AI requirements are dispersed into different components of the application, the AI infrastructures for ridesharing are built to meet each teams' demands. As we can observe from the open-source infrastructures, a flexible software layer is built to facilitate fast CI/CD of machine learning models, which are developed on top of CPU/GPU clusters.

3.11.4 Key Market Opportunities

The infrastructure for ADAS and other AI applications can be provisioned either from public clouds or on-prem clouds. Many public clouds provide ADAS services. For example, AWS has ADAS and self-driving services, which attracted many customers. The benefit of a cloud-based ADAS service is that it will be ready to use in little time, and the cloud can satisfy many enterprise requirements from geo-replication to GDPR requirements. Many autonomous driving development companies are using cloud services to reduce the go-to-market time.

However, one disadvantage is that the public cloud infrastructure is not built to meet the exact needs of the enterprise, particularly when the enterprise uses proprietary technologies for its vehicle. As an example, Tesla has its on-prem training facility and applies unique methodologies for its self-driving vehicles. In other words, public clouds provide a common service for all customers. If a company needs to break out from the autonomous driving crowd, it may consider building its proprietary systems. On-prem clouds would be a good fit for this purpose.

Another consideration is cost. Public cloud has cost savings when the scale is small – many customers shared the initial cost. It makes perfect sense to use public clouds when the company is small and growing rapidly. But when the business is stabilized and a large data center is needed, self-built data centers and on-prem clouds may have a lower cost because the large-scale data centers can have the same savings as public clouds due to the scale.

Many companies have a hybrid cloud strategy. For example, Uber has used both public clouds and on-prem clouds, a strategy referred to as "tripod strategy", which "combines the use of public cloud services with standardized on-premise server racks in its colocation facilities, all separately handling compute, storage, database, and GPU workloads" [25] . In this article, Dean Nelson, Uber's head of compute, recognized the benefits of building on-prem clouds using "building blocks" referred to as Uber Metal [25].

3.12 AGRICULTURE

3.12.1 Sector introduction

Agriculture is one of the oldest and most important professions in the world. Throughout human history, whenever there is a new technology, people will try to apply it to agriculture, and there is no exception to AI. Currently, there are limited direct applications of AI technology in agriculture, however, life science and genetic research have a huge impact on agriculture, which can be considered as an indirect impact. Since we already covered life science in a separate section, we will only focus on direct AI applications in this section [26].

3.12.2 Key players

Key players include Farmers, Granular, John Deere, BlueRiver, AgroStar, and FarmLogs.

3.12.3 AI applications for business

Here are a few popular applications of AI technology in agriculture:

- **Helping analyze farm data**
  AI can help analyze a variety of things in real-time such as weather conditions, temperature, water usage, or soil conditions collected from their farm to better inform their decisions. For example, AI technologies help farmers optimize planning to generate more bountiful yields by determining crop choices, the best hybrid seed choices, and resource utilization.
- **Help to improve harvest quality and accuracy (precision agriculture)**
  Precision agriculture uses AI technology to aid in detecting diseases in plants, pests, and poor plant nutrition on farms. AI sensors can detect and target weeds and then decide which herbicides to apply within the right buffer zone. This helps to prevent the over-application of herbicides and excessive toxins that find their way into our food.
- **Seasonal forecasting models to improve productivity**
  AI can be used to produce some seasonal forecast models that can predict upcoming weather patterns months to assist decisions of farmers. This can be very valuable to farmers.
- **Computer vision and deep learning enhanced farm monitoring**
  From drones, AI-enabled cameras can capture images of the entire farm and analyze the images in near-real-time to identify problem areas and potential improvements. Unmanned drones can easily cover far more land in much less time than humans on foot allowing for large farms to be monitored more frequently.
- **Indirect impact: robotic labor**
  With the modern technology boom and rich options for young people, there are fewer and fewer new labor forces entering the agriculture market. Luckily many labor requirements on farms can be replaced or heavily-reduced by a combination of smarter and more efficient farming machines, or even fully AI-capable robots.

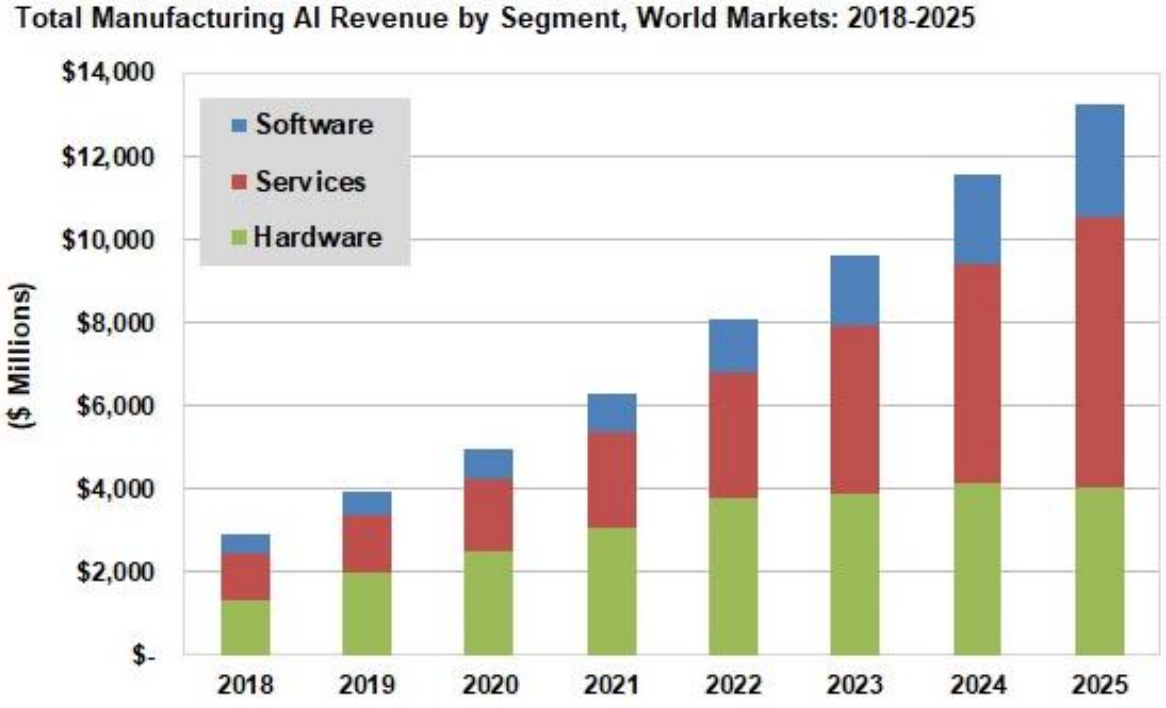
3.12.4 Key Market Opportunities

Although agriculture is one area that AI could help, it is heavily relying on simple rule-based AI or indirectly benefit from other industries such as genetic research and robotics. Combined with the lean revenue from the farming industry, we do not believe there is a substantial business opportunity there.

3.13 INDUSTRIAL

3.13.1 Sector introduction

The manufacturing industry makes up a crucial part of the whole economy, and they constantly face many challenges throughout their manufacturing cycle, such as industrial design, production automation, maintenance, and troubleshooting when things go whacky. There are many fields where AI and machine learning can help improve efficiency and reduce cost. The chart below shows solid growth in AI revenue in the manufacturing sector expected in the coming years.



3.13.2 Key players

Key players include Siemens, Schneider, Avera, Predix, Uptake, and Kayrros.

3.13.3 AI applications for business

USM listed the top 10 use cases for AI in the industrial sector:

#1 Quality Checks
#2 Predicts Equipment Failure
#3 Equipment Predictive Maintenance
#4 Digital Twins (Sensor data-based digital representation of physical objects)
#5 Supply-Chain Management
#6 Forecast Product Demand
#7 Inventory Management
#8 Price Forecasts
#9 Robotics in Manufacturing
#10 Customer Management

These are the most valuable area that already sees major benefits from AI. They mainly help in improving manufacturing efficiency and precision, while reducing cost.

3.13.4 Key Market Opportunities

Manufacturing companies are typically not tech-savvy, building and maintaining a private data center is normally not in their priority even for very large manufacturing companies. Solutions carved for industrial

companies need to be as easy to use as possible, with the highest level of integration, most convenient management interface. There should be plenty of edge nodes or remote station type compute and storage needed in the industrial environment, with the ability to closely work with a central database or cloud. We can consider a dHCI cluster with GPU/NPU enabled compute node and low-mid range storage as a recommended solution for most small industrial customers, or a remote station for larger customers. The whole cluster should be contained within 1 rack, pre-loaded, and preferably managed through a single management interface. In case a large customer is interested in building and maintaining their own central data center, a traditional compute and storage solution should be sufficient.

3.14 OTHERS

There are other industries such as semiconductors, oil and gas, and telecommunications. They follow a similar pattern as other industries.

4 AI WORKLOAD CLASSIFICATION

One of the major requirements from AI/ML infrastructure providers is how to systematically categorize computing and data transfer demands for AI/MI workloads. Different AI/ML applications put pressure on different parts of the infrastructure.

Based on related statistics, around 62% of total execution time among AI/ML workloads on average is spent in weight and gradient communications. 60% of workloads can be potentially sped up by using AllReduce architecture exploiting high-speed links between GPU interconnects. And 1.7x speedup can be achieved when Ethernet bandwidth is upgraded from 25G to 100Gbps.

AI/ML has rather different phases: ML analytics and big data processing, training, and inference. Each phase has rather different workload characterizations. The table below summarizes high-level storage requirements in a different phase of AI workflow:

	DATA CHARACTERISTICS	STORAGE REQUIREMENT	NETWORK
INGEST	write-heavy, sequential I/O, mixed file sizes	high throughput	10-100GE
PREPARE	read/write heavy, random/sequential I/O, mixed file sizes	random I/O performance	10-100GE
TRAIN	read/write heavy random I/O, small files	highly parallel, high bandwidth (GBs/s), low latency (<1ms)	100GE/IB RDMA preferred
INFERENCE	read-heavy, multi-tenant	extreme low latency	100GE RoCE / IB

4.1 ML ANALYTICS AND BIG DATA PROCESSING

This phase is commonly referred to as the data cleaning and analytics phase, equivalent to “ingest” and “prepare” in the summary table above. It is more related to big data and analytics than ML. This phase can be broken down into several stages: Data ingestion, Data cleaning, and data egress.

The data ingestion phase involves a massive amount of data streaming into a data lake. This is a typical large quantity of files with sequential write workloads.

Data Cleaning plays an important role in the field of Data Management as well as Analytics and Machine Learning. Data cleaning typically demands large or small sequential read/write workloads with a large throughput.

4.2 TRAINING

The most important factor to support training performance is to NOT let other infrastructure components be bottlenecks. Let GPU be! That is easier said than done. Training workloads are normally more I/O intensive. Latency, throughput as well as parallel access are all requirements for storage and network. Random small file read with large quantities is a typical workload for training (image processing, computer vision, etc.). Individual files seldom big enough to deserve any attention, but there could be millions of files at scale. Total aggregate throughput can also be up to more than multiple GB/S. Due to this requirement, a lot of IT infrastructure for AI/ML likes to deploy a parallel file system for AI storage needs. A lot of new technologies like GDS and RDMA are leveraged to speed up data transfer between storage and GPUs.

The outputs of the AI/ML training phase are more easily manageable for IT infrastructure. They are often small enough that there is no issue with modern enterprise IT systems.

4.3 INFERENCE

During the inference phase, AI/ML algorithms apply pre-trained models on incoming data to make a decision/prediction. The amount of data involved in this phase is typically small, however, the decision normally needs to be made as soon as possible, and requests could come frequently. This operation requirement brings a different challenge to the storage system compared to the data processing and training phase. The capacity of the storage systems is not key consideration anymore and read I/O latency has become the number one important factor. For small-scale systems, a local NVME SSD will do the job, assuming the training model is actively synced up with the training side of the pipeline. NVME over RDMA fabric is most suitable for large-scale or distributed systems.

5 BENCHMARK

There are many popular benchmark tools and platforms in the industry, these generic results provide valuable information as a reference for AI and storage applications. We will go through a few most popular and widely accepted benchmarks in the storage industry in this chapter.

5.1 STORAGE BENCHMARK

5.1.1 Primary Storage Benchmark

Primary storage is very sensitive to both latency and throughput in block access. One of the most widely accepted primary storage benchmark scoreboards is SPC-1 [27]. Huawei OceanStor Dorado is the number 1 top performer on this list, with a single cluster offering 21 million IOPS throughput and as low as 0.05ms latency on block access.

Here is the latest published result:

Rank	Performance (SPC-1 IOPS)	Test Sponsor	Price-Performance (per SPC-1 KIOPS™)	Submission Identifier	Tested Storage Product
#1	21,002,561	HUAWEI	CN¥ 2913.78	A32018	OceanStor Dorado 18000 V6
#2	11,000,576	macrosan	$ 385.64	A32020	MS7000G2-Mach
#3	10,001,522	FUJITSU	$ 644.16	A32009	ETERNUS DX8900 S4
#4	7,520,358	inspur 浪潮	$ 386.50	A32014	Inspur AS5600G2
#5	7,000,565	HUAWEI	$ 376.96	A31017	Huawei OceanStor Dorado18000 V3

5.1.2 File System Benchmark

NAS is a very popular format of storage in the AI/big data area. One of the most popular file storage benchmarks is IO 500 by Virtual Institute [28]. This list covers a range of popular file systems from industry and academy,

such as file system type Lustre and DAOS; vendor list includes Intel, WekaIO, and Argonne National Lab. Huawei currently does not have a presence on this list. The figure below shows top entries on the current IO500 list:

#	information								io500		
	list id	institution	system	storage vendor	filesystem type	client nodes	client total procs	data	score	bw	md
										GiB/s	kIOP/s
1	sc20	Pengcheng Laboratory	Pengcheng Cloudbrain-II on Atlas 900	Pengcheng Laboratory	MadFS	255	18360	zip	7043.99	1475.75	33622.19
2	isc20	Intel	Wolf	Intel	DAOS	52	1664	zip	1792.98	371.67	8649.57
3	sc19	WekaIO	WekaIO on AWS	WekaIO	WekaIO Matrix	345	8625	zip	938.95	174.74	5045.33
4	isc20	TACC	Frontera	Intel	DAOS	60	1440	zip	763.80	78.31	7449.56
5	isc20	Argonne National Laboratory	Presque	Argonne National Laboratory	DAOS	16	544	zip	537.31	108.19	2668.57
6	sc19	National Supercomputing Center in Changsha	Tianhe-2E	National University of Defense Technology	Lustre	480	5280	zip	453.68	209.43	982.78
7	sc20	Intel	Endeavour	Intel	DAOS	10	640	zip	353.72	43.59	2870.64
8	isc20	KISTI	NURION	DDN	IME	2048	2048	zip	282.45	515.59	154.74
9	isc20	Oracle Cloud Infrastructure	BeeGFS on Oracle Cloud	Oracle Cloud Infrastructure	BeeGFS	270	3240	zip	267.25	293.05	243.73
10	sc20	JCAHPC	Oakforest-PACS	DDN	IME	2048	4096	zip	253.57	697.20	92.22
11	sc19	NVIDIA	DGX-2H SuperPOD	DDN	Lustre	10	400	zip	249.50	86.97	715.76
12	sc20	EPCC	NextGENIO	BSC & JGU	GekkoFS	10	3800	zip	239.37	45.79	1251.32
13	sc19	University of Cambridge	Data Accelerator	Dell EMC	Lustre	128	2048	zip	229.45	131.25	401.13

5.1.3 Database Benchmark

Major database vendors frequently offer their benchmarking utility to evaluate storage system performance. Some are only available to customers and partners, such as SAP HANA standard application benchmark tool. There are also some open tools such as Microsoft Diskspd and SQLIO. Due to the vast difference between different database software, there is not any meaningful industry-wide scoreboard for storage performance in database applications. Each application tends to maintain its list or leave it up to the customer to decide.

5.1.4 Generic I/O Benchmark Tools

There are plenty of generic storage I/O performance benchmark tools available, the most popular ones for testing enterprise storage arrays include iometer and fio. The fio community has many scripts available for simulating all kinds of common types of workload such as database, big data analytics, etc.

Fio has comprehensive documentation available online [29].

The project source code is hosted on GitHub [30], with plenty of examples showing how you can use a script [31] to simulate all kinds of workload to match your real application.

5.1.5 GPUDirect Storage Benchmark

In 2020, Nvidia published a new way for GPU to directly access storage through RDMA fabric without involving CPU and system memory in the middle. This greatly improved high-end AI-storage cluster overall performance. The GPUDirect software suite includes a benchmark tool “gdsio” by Nvidia to test I/O performance of the cluster.

Here is a sample command for benchmarking local NVME disk direct access through GDS:

```
/usr/local/cuda/gds/tools# ./gdsio -f /mnt/test/testfile1 -d 0 -w 4 -s 10G -i 1M -I 0 -x 0
IoType: READ XferType: GPUD Threads: 4 DataSetSize: 10212352/10485760(KiB) IOSize: 1024(KiB) Throughput:
2.681944 GiB/sec, Avg_Latency: 1456.483233 usecs ops: 9973 total_time 3.631417 secs
```

Simillarly, GDS can directly access remote NVME over fabric using NVMEoF protocol:

```
/usr/local/cuda/gds/tools# ./gdsio -f /nofmnt/testnof -d 0 -w 4 -s 10G -i 1M -I 0 -x 0
IoType: READ XferType: GPUD Threads: 4 DataSetSize: 1016832/1048576(KiB) IOSize: 1024(KiB) Throughput:
2.236000 GiB/sec, Avg_Latency: 1743.542703 usecs ops: 993 total_time 0.433688 secs
```

See the GDS website [32] for more details on how to use the utility to troubleshoot and benchmark GPUDirect storage.

5.2 AI/ML Compute Benchmark

Due to the complexity and huge variety of AI/ML applications, there is a lack of industry-standard benchmark tools or widely accepted scoreboards. One commonly accepted standard is the raw computing power in terms of how many TFLOPS the system can handle, this does not necessarily have a linear correlation with different applications, but still provide a good reference point.

Since the AI hardware market is mostly dominated by Nvidia, a very good reference for checking hardware raw computing power is the wiki page for Nvidia GPU [33]. It contains a detailed specs comparison table of GPUs within each generation/family. For example here is the table for Tesla data center GPU [34], here is a sub-section of the Tesla table with the most important information for AI compute efficiency.

Model	Micro-architecture	Chips	Shaders	Memory				
			Cuda cores (total)	Bus width (bit)	Size (GB)	Bandwidth (GB/s)	Single precision (MAD or FMA)	Double precision (FMA)
P100 GPU accelerator				3072	12	549	8071–9340	4036–4670
V100 GPU accelerator (mezzanine)	Volta	1× GV100-895-A1	5120	4096	16 or 32	900	14899	7450
V100 GPU accelerator (PCIe card)		1× GV100					14028	7014
T4 GPU accelerator (PCIe card)	Turing	1× TU104-895-A1	2560	256	16	320	8100	Unknown
A100 GPU accelerator (PCIe card)	Ampere	1× GA100-883AA-A1	6912	5120	40	1555	19500	9700

Some generic tool suite such as ai-benchmark [35] offers a wide range of coverage on popular algorithms. It can be used as a good reference too. Here is the top part of a detailed list of GPU AI benchmark score ranking provided by ai-benchmark.

Model	TF Version	Cores	Frequency, GHz	Acceleration	Platform	RAM, GB	Year	Inference Score	Training Score	AI-Score
Tesla V100 SXM2 32Gb	2.1.0	5120 (CUDA)	1.29 / 1.53	CUDA 10.1	Debian 10	32	2018	17761	18030	35791
Tesla V100 SXM2 16Gb	2.1.0	5120 (CUDA)	1.31 / 1.53	CUDA 10.1	Red Hat 7.5	16	2017	17251	17836	35086
Tesla V100 PCIE 32Gb	2.1.0	5120 (CUDA)	1.23 / 1.38	CUDA 10.1	Debian 10	32	2018	16530	17865	34394
Tesla V100 PCIE 16Gb	2.1.0	5120 (CUDA)	1.25 / 1.38	CUDA 10.1	Red Hat 7.5	16	2017	16511	17837	34347
NVIDIA Quadro GV100	1.14.0	5120 (CUDA)	1.13 / 1.63	CUDA 10	Debian 10	32	2018	16748	17132	33880
NVIDIA TITAN V	2.1.0	5120 (CUDA)	1.20 / 1.46	CUDA 10.1	Ubuntu 18.04	12	2017	16192	17215	33406
NVIDIA TITAN RTX	2.1.0	4608 (CUDA)	1.35 / 1.77	CUDA 10.1	Ubuntu 18.04	24	2018	16084	17255	33339
GeForce RTX 2080 Ti	2.1.0	4352 (CUDA)	1.35 / 1.55	CUDA 10	Debian 10	11	2018	16042	16828	32870
NVIDIA Quadro RTX 8000	2.1.0	4608 (CUDA)	1.40 / 1.77	CUDA 10.1	Debian 10	48	2018	13014	14637	27651
NVIDIA Quadro GP100	2.0.0	3584 (CUDA)	1.30 / 1.44	CUDA 10	Red Hat 7.4	16	2016	12264	13436	25700
NVIDIA TITAN Xp	2.1.0	3840 (CUDA)	1.41 / 1.58	CUDA 10.2	Debian 10	12	2017	11948	12922	24870
GeForce GTX 1080 Ti	2.1.0	3584 (CUDA)	1.58 / 1.60	CUDA 10.2	Debian 10	11	2017	11914	12473	24386
GeForce RTX 2080 SUPER	2.1.0	3072 (CUDA)	1.65 / 1.82	CUDA 10.1	Windows 10	8	2019	11513	12734	24247
GeForce RTX 2070 SUPER	2.1.0	2560 (CUDA)	1.61 / 1.77	CUDA 10.2	Ubuntu 18.04	8	2019	11472	12710	24182

Tesla V100 is the GPU that is used in the original version of Nvidia DGX, the latest generation DGX-2 uses a more advanced Tesla A100 card, which is even more powerful.

6 Reference Architecture

6.1 Summary

AI/ML especially deep learning has made significant progress during the last 10 years, particularly in the fields of NLP, computer imaging, vision, and autonomous driving. It is powered and accelerated by advanced research of new algorithms and large datasets with higher than ever computational powers.

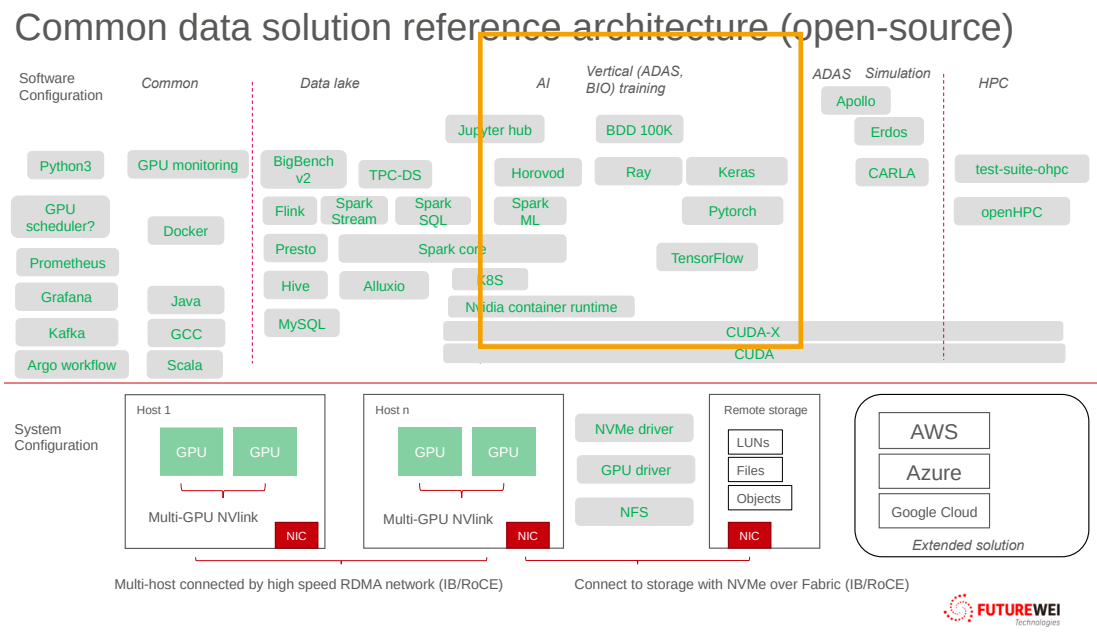
To fully utilize the full potential of ever-increasing computational power for ML and reduce overall costs of ML, a new infrastructure architecture shall be designed to accommodate those new workloads, reduce overall complexities of infrastructure and strike a balance among computing, network, and storage. This section is trying to classify different choices for overall architecture and give recommendations to build computation, memory, storage, and network as well as software ecosystem.

Based on application requirements for performance, storage, and network, we break down AI/ML configuration into 3 typical tiers:

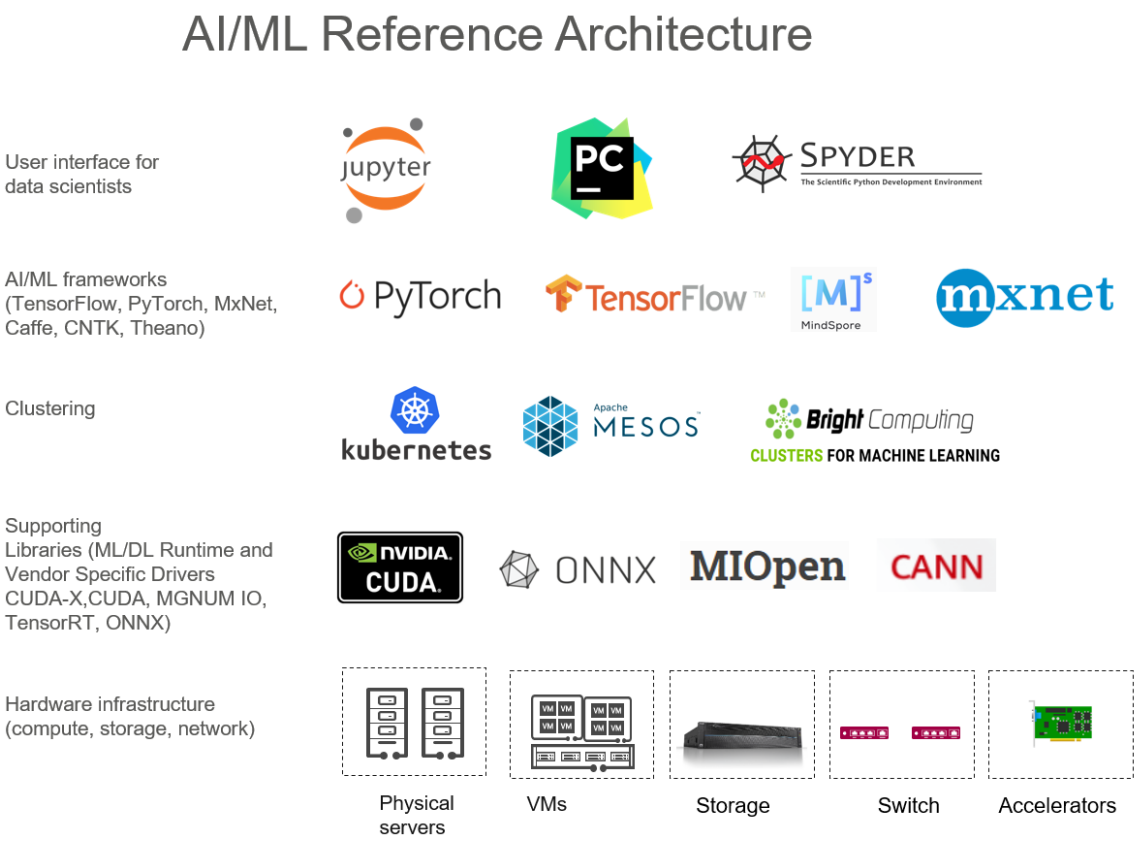
6.2 AI/ML Reference Architecture

Regardless of different tiers, there is a common need for an AI/ML architecture as shown below. Here is a common data solution reference architecture for HPC and AI/ML. As you will know, it is rather complex and overwhelming for data science to choose the right hardware, software, and frontend to start their works. The

goal is to not let data scientists worry about those configurations among storage, network, and GPU. Rather let them focus on data science only.



By consolidating different use cases, requirements, and other factors, we propose the following reference architecture. We will describe each sector in detail:



6.2.1 AI/ML Data Science Front End

An AI/ML data science front end allows data scientists to keep track of errors and maintain clean code. It is a one-stop-shop of IDE environment for AI and ML. Jupyter notebook, Jupyter lab, Spyder, Guleviz, Orange, RSstudio, VSC Code are some of the typical front-ends with the Jupyter series being the most popular ones. The front end should allow end users to choose the framework to use and which cluster to run the training program. The front end should mask the gory details of the underneath layers and expose the necessary choices to the end users.

6.2.2 AI/ML frameworks

In this layer, an end user can choose from many frameworks. The most popular two today are Tensorflow and PyTorch. Many others have their own user communities. Many vendors push their own AI/ML library to influence the community. This layer is generally open-source based.

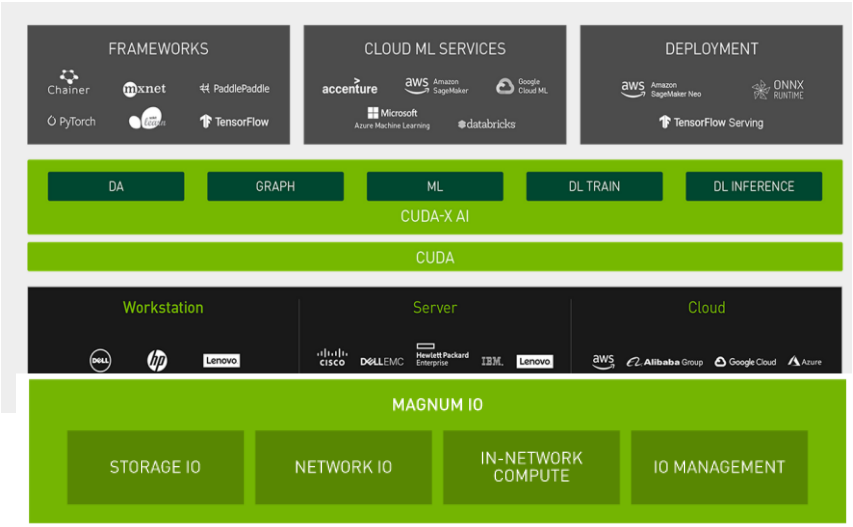
6.2.3 Cluster Resource Management

A cluster resource management layer is to manage different compute cluster resources to provide upper lay resource isolation and provisioning. It also facilitates customer application deployments on top of resource management and orchestration software (Yarn, K8S, Mesos, etc.). The most popular cluster resource management software is Jupyter enterprise gateway and its commercial enhanced variant – Bright Cluster Management for AI/ML.

6.2.4 Supporting libraries

This section is hardware related because each vendor has its own libraries based on the special hardware. For example, CUDA for Nvidia GPUs and MIOpen for AMD Instinct GPUs. There are many AI/ML solutions that customers can choose from. Being at top of the game, Nvidia provides the most complete ecosystem in this area with supporting most major frameworks and hardware optimizations.

The following diagram shows an optimized stack for Nvidia AI/ML ecosystem:



Source : https://developer.nvidia.com/blog/accelerating-io-in-the-modern-data-center-magnum-io-architecture/

Telemetry and troubleshooting across compute, network, and storage layers.

Cumulus NetQ, Mellanox UFM

The GPU bypasses the CPU and system memory, and accesses remote storage via 8X 200 Gb/s NICs, achieving up to 1.6Terabits/s of raw storage bandwidth.

GPUDirect, Mellanox NVMe SNAP

NVIDIA NVLink® fabric and RDMA-based network IO acceleration reduces IO overhead, bypassing the CPU and enabling direct GPU to GPU data transfers at line rates

DPDK, GPUDirect RDMA, HPC-X, NCCL, NVSHEM, UCX, ASAP

Offloading to “network processors”.

Bluefield DPU, MPI tag matching, Mellanox SHARP

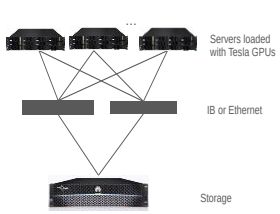
FUTUREWEI Technologies

6.2.5 Infrastructure

This is a fundamental building block of ML/AI, all the performances and computation power come from this layer. Generally, these layers include 3 major parts: Compute (including GPUs), storage, and network. Compute usually is built with high GPU density high-performance servers with larger memory and faster local NVMe drives. A good example is the Nvidia DGX system. Fast speed and big throughput with RDMA capability network shall also be provided to satisfy the bandwidth requirement of GPU and applications. For large-scale training, a scale-out NAS storage system shall be provided.

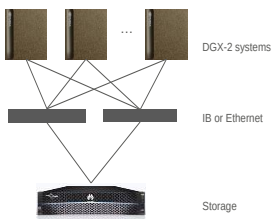
Below are shown 2 typical configurations for this type of infrastructure:

Target #1: GPU-heavy cluster (DGX-2 like) storage reference architecture (less dependent, less expensive)



A quick assembly of the components will yield a rack loaded with GPU-heavy servers and storage, connected by RoCE or IB network.

Target #2 (Stretch goal): DGX-2 Storage reference architecture (Depending on the collaboration with Nvidia and budget)



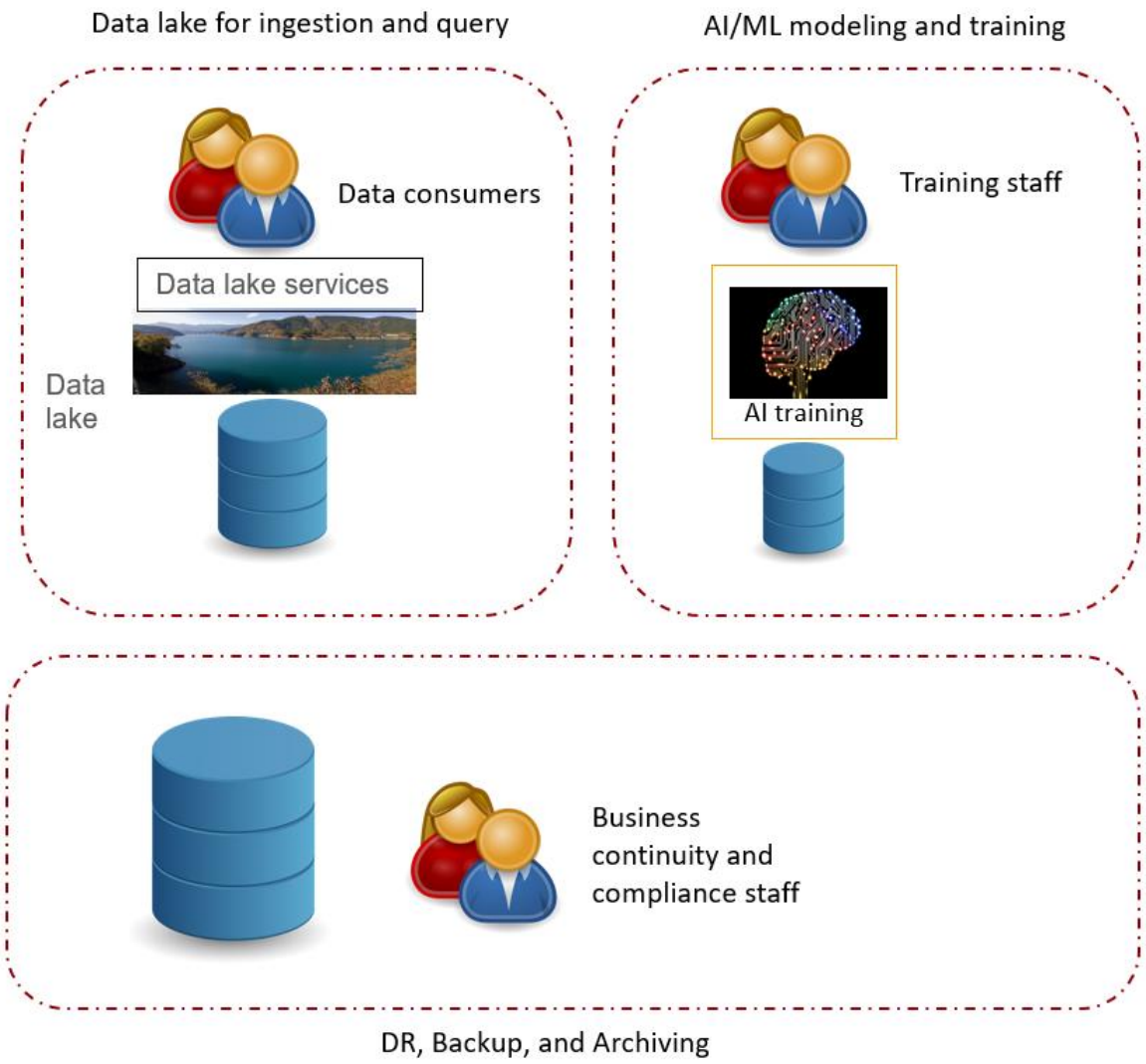
A quick assembly of the components will yield a rack loaded with two DGX-2 systems.

- *Compute*
  In this reference architecture, the compute can be composed of any mainstream CPUs, such as Intel Xeon, AMD, or various ARM-based CPUs. The GPU/TPU can also use different vendors. As Nvidia is having a strong position in the GPU world, an Nvidia GPU accelerated system is a popular choice. But other chips are possible if they support major AI/ML frameworks, such as TensorFlow and Pytorch.

  The compute-storage disaggregated architecture allows scale-out of the computing resources and storage resources independently. Therefore, it is well suited for large-scale deployment. For smaller environments, a distributed hyper-converged (dHCI) environment can be deployed. The dHCI solution has both the advantage of the traditional HCI (fast deployment, easy management) and the advantage of the disaggregated architecture (flexibility).

- *Network*
  The network section can use multiple high-speed networks. Higher bandwidth and lower latency are critical to the performance of this disaggregated architecture. However, due to cost reasons, users may not want to use the fastest network. RDMA over fabric and RDMA over Ethernet (RoCE) are the two most popular network standards used in this reference architecture. Users may choose a network that fits the budget and bandwidth needs.

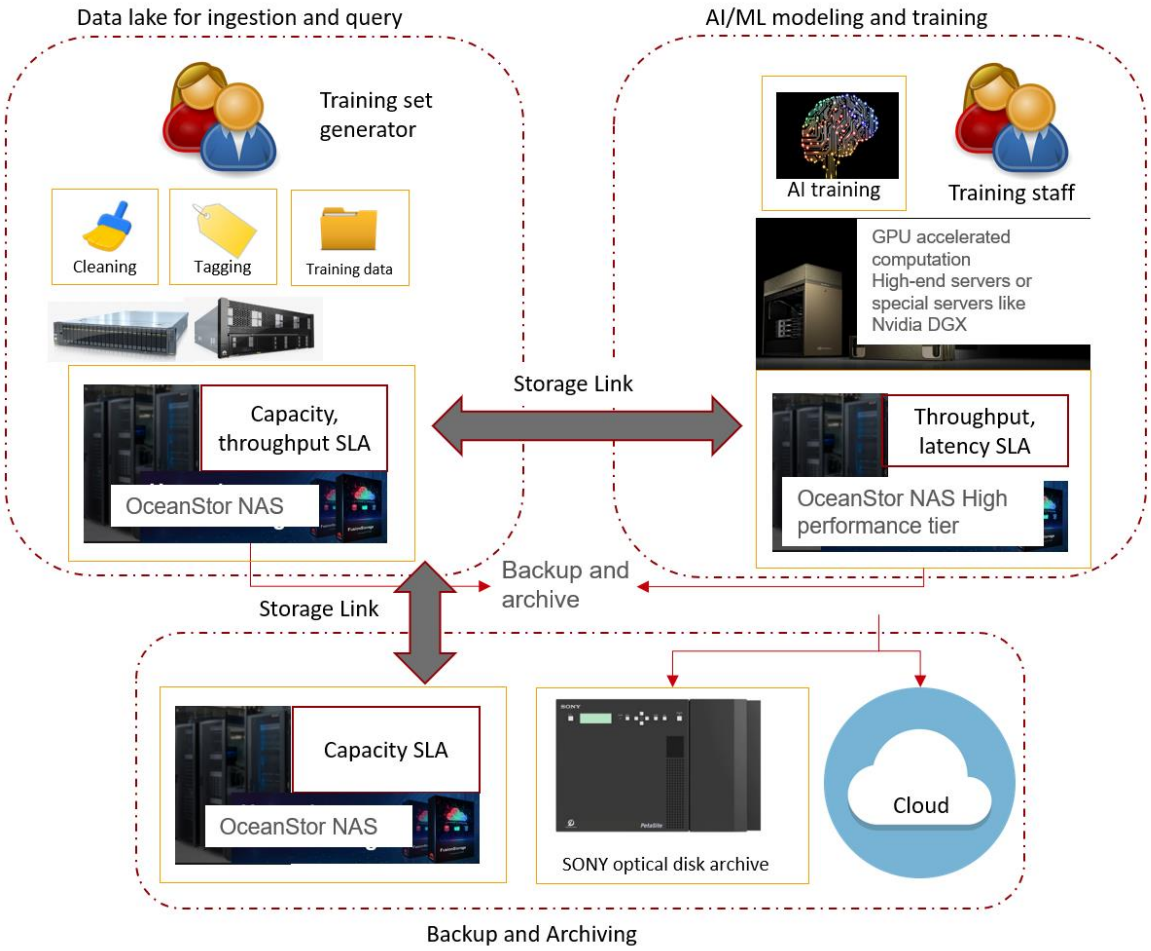
- *Storage*
  There are three areas where storage requirements arise. They are the data lake, AI/ML training, and DR/backup/archiving. These three areas have different focuses, even though they have overlapping goals. The data lake users may have stricter criteria of the unit price of each GB, due to the amount of data needed. But they also put performance and throughput into consideration. The AI/ML training users are concerned with the speed of training and the performance of underlying storage. The DR/backup/archiving users may have an even bigger capacity challenge because multi-year data will be kept for compliance with regulations. But latency is generally not an issue.



6.3 HIGH-END AI CLUSTER

For mission-critical AI applications that require ASAP response and not very sensitive to cost, the solution should include the following components:

- Compute: A powerful computing node such as Nvidia DGX-2 or equivalent servers.
- Network: RDMA fabric (100Gb RoCE or InfiniBand) is recommended for connection between the compute node and storage system, so the AI engine can access data through NVMeOF with minimal latency and high throughput.
- Storage: The storage in the solution can be mid to high-end all-flash array depends on the estimated workload and capacity requirements of the targeted application.



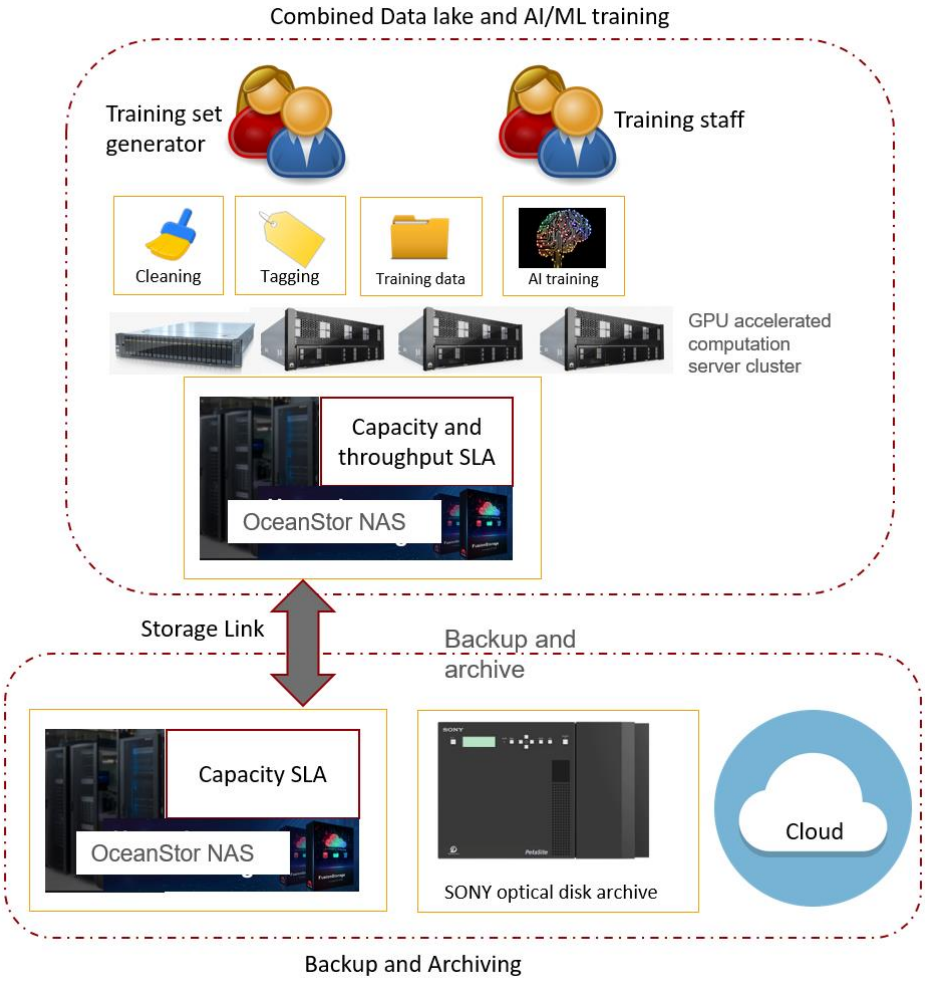
This solution is suitable for tier 0 applications such as hedge funds, where each second delay in making a correct decision may mean millions of dollars. Other use cases include the insurance industry where clients could be waiting online for a quote, and credit card company fraud detection.

6.4 Mid-range AI Cluster

Many AI application clients are seeking a balance between performance and cost, most of them can be classified into this solution group. The proposed solution should maximize performance per dollar spent while satisfying the application's minimal requirements. Here is an example of components for a mid-range AI compute / storage cluster:

- Compute: Multi-GPU/TPU server cluster.
- Network: 25/40Gbps RoCE RDMA fabric is recommended for connection between the compute node and storage system. This setup will allow NVMeOF access for best performance and cost less than the 100Gbps solution.

- Storage: Mid-range all-flash array with enough capacity and front-end I/O modules to satisfy minimum application requirements.

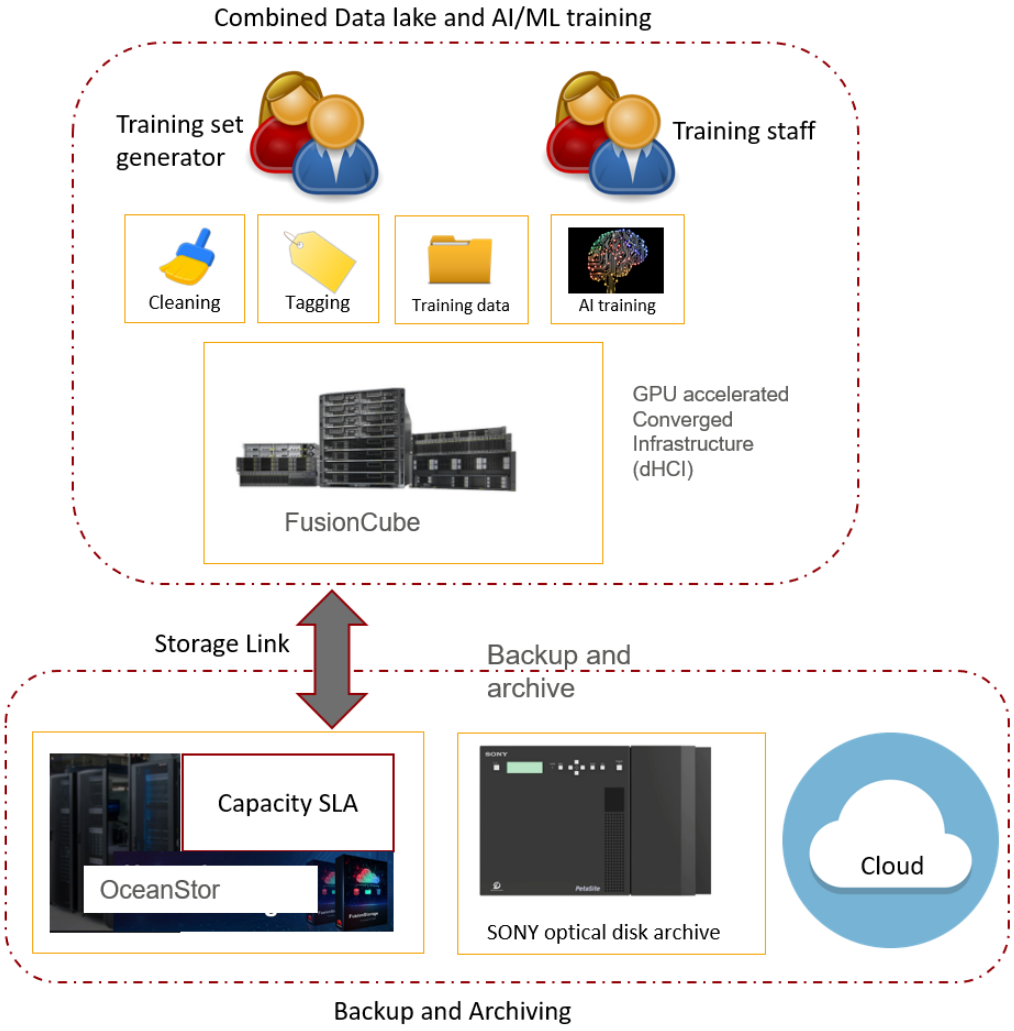


A wide range of applications in enterprise, industry, and academy can use this solution, such as hospitals for AI-assisted diagnosis, advertising companies for offline analysis, automotive industry for ADAS training, etc. See chapters 3 and 4 for more ideas.

6.5 Entry-level AI Cluster / Remote Office

Some AI applications do not need very large scale to work, or by nature deployed in a distributed way with multiple remote sites that each needs to be able to perform AI tasks on their own. For such kinds of small-scale AI applications, flexibility, convenience, and low cost are frequently the most important factors when clients choose their hardware solution. HCI or dHCI solutions should be very good candidates for such scenarios. For example, in the case of HCI, a small HCI cluster with the following characteristics can be a very good candidate:

- 2-3 nodes redundant HCI cluster taking half-rack or shipped as a pre-configured box
- Each node contains 1-4 GPU/TPU
- Key enterprise-grade storage features built-in the HCI stack, such as snapshot, clone
- Inter-node communication uses 10/25Gbps ethernet with optional RoCE configuration
- Integrated management interface for storage and compute



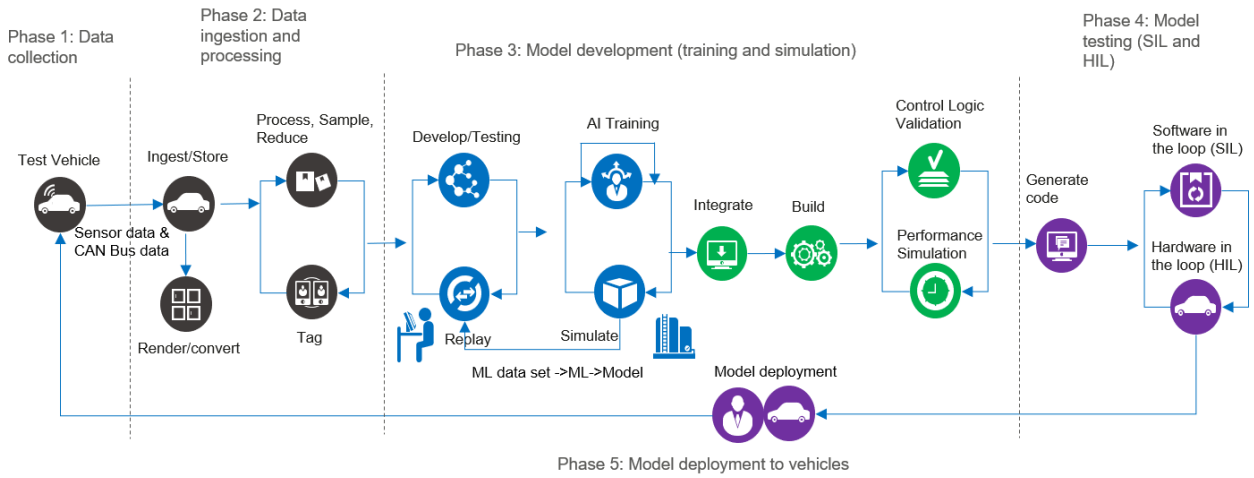
Example of such scenario includes a remote office for financial and insurance companies, life science research labs, medical institute field offices, etc. See chapters 3 and 4 for more market opportunity ideas.

7 SELECTED USE CASES

7.1 ADAS

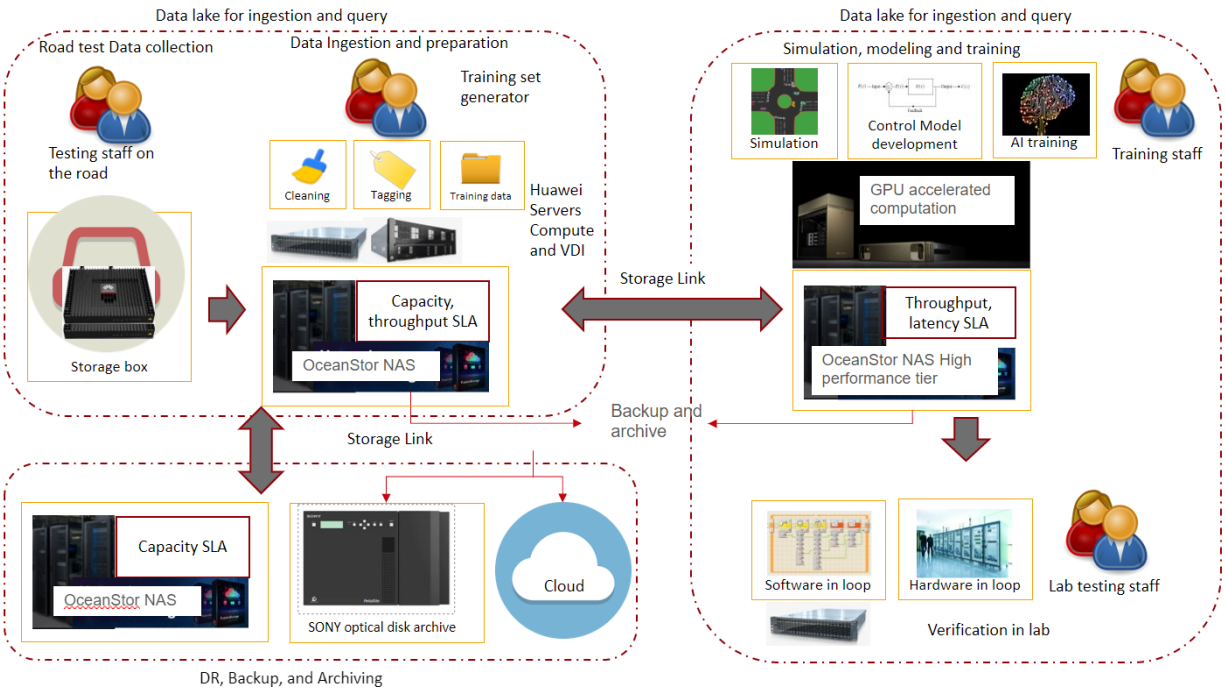
The ADAS development process can be divided into the following phases:

1. Data collection
2. Data ingestion and processing
3. Model development with training and simulation
4. Model testing (SIL and HIL)
5. Model deployment to vehicles



First, data is collected through a fleet of vehicles traveling on target terrain. Data may include time sequence data from sensors, cameras, lidar, and positioning systems. The estimated size of data may reach 60TB per car each day. Second, the collected data is processed to have a clean format and tagged information for machine learning usage. Today the tagging process can be automated but there is still some work left for humans. The third phase, model development with training and simulation, is the core piece of the ADAS infrastructure. The training set is fed into neural networks to detect objects, project trajectories, and plan actions of the vehicle. Computers can use real road data and simulated road data to train the model until the model reaches a satisfactory goal. The companies that can train the model fast and accurately will provide better products and shorter go-to-market turnaround times. In the fourth phase, the trained model will be deployed to labs to have software-in-the-loop (SIL) and hardware-in-the-loop (HIL) testing. SIL testing is performed in a software-based environment, whereas HIL testing emulates the real car model for control systems. In the fifth phase and the final phase, the polished model is deployed on the autonomous vehicle and it can generate more data in future road tests.

Each phase of the process has a different requirement for storage. It is well known that each enterprise has different goals so that the sizing of the capacity could be very different among different car manufacturers. The more data acquired, the more demanding the system will become.



Phase 1 needs on-vehicle equipment to store data from all sensors. There are other requirements for shock resistance and temperature.

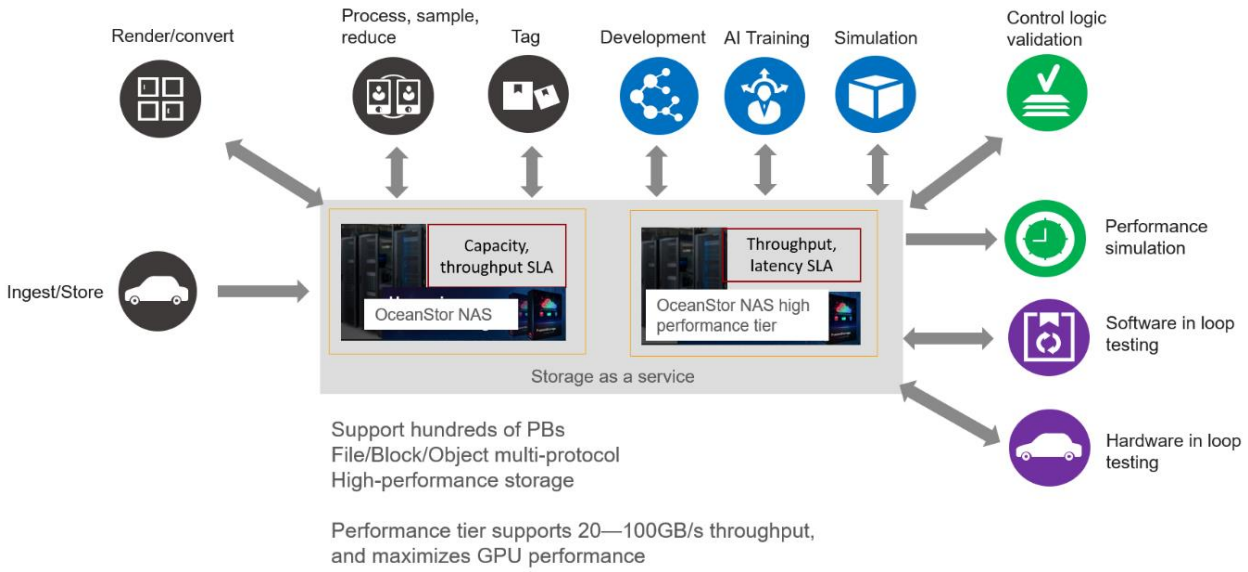
Phase 2 needs a big data processing system, which can be hundreds of PetaBytes. The actual capacity depends on the enterprise's business goals. The reference architecture defines such a system.

Phase 3 needs a training system that can handle millions of pictures, which can go into PetaBytes. Depending on the requirement of training time, the bandwidth of reading can be high. Nvidia GDS is designed to provide high bandwidth to GPUs.

Phase 4 requires the SIL system to verify a large part of the scenario data (e.g., 30%) and write results to the storage system every day. All of these require a large bandwidth system. A high-performance storage cluster is needed to handle this workload.

Phase 5 does not have specific requirements for storage.

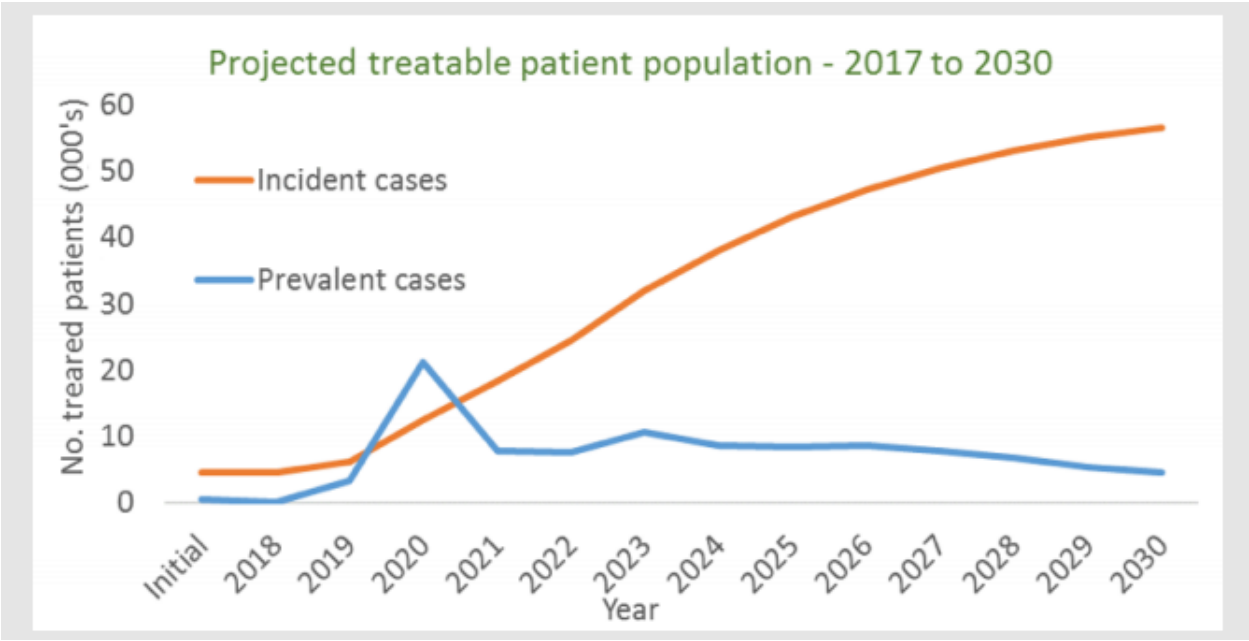
The storage service can be divided into two tiers. One tier is capacity sensitive, and the other tier is performance sensitive, particularly for the AI development environment. Details of each environment are based on customer requirements. One requirement is that the system should be scalable as the environment changes.



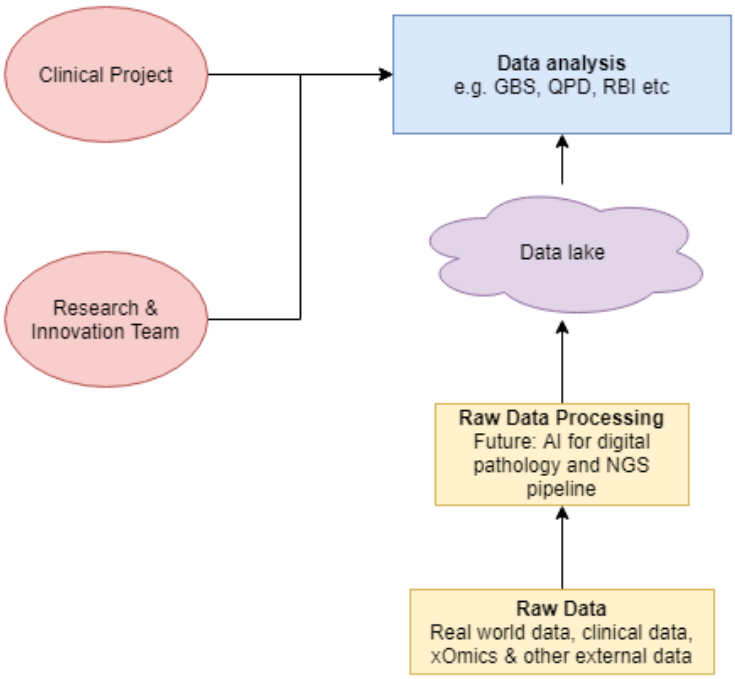
The previous diagram shows an environment using the OceanStor storage platform. The storage as a service layer provides a foundation for persistent storage. The reference architecture can be used to implement the phases.

7.2 HEALTHCARE AND LIFE SCIENCE

AI found extensive use in modern medical and life science, throughout a wide range of areas such as drug discovery, disease study, diagnoses, etc. The next 10 years will likely see significant changes in the US healthcare system with major improvements in the treatment paradigms for numerous diseases that previously had high morbidity and mortality. These will require the overall system to adapt, particularly in terms of how treatments are reimbursed and financed, as we move from chronic palliative therapies to acute curative ones. Below is predicted genetic treatment market growth [36]:



The picture below shows a typical life science research workflow.



There are a few main pain points in the life science AI applications, the first is capacity.

Genetic research had multiple breakthroughs in recent years and has taken off in medical applications such as cancer diagnosis and treatment. The genetic applications market is forecasted to grow explosively in the coming years.

Quoting data from strand-NGS v2.9 (A popular genetic analysis software), here is the computing and storage requirements for analyzing a single genome:

The storage requirement for a single human genome:

	Coverage	No. of Reads	Read Length	BAM File Size	Strand NGS Size
Whole Genome	37.7x	975,000,000	115	82 GB	104 GB
Whole Genome	38.4x	3,200,000,000	36	138 GB	193 GB
Exome	40x	110,000,000	75	5.7 GB	7.1 GB

The 1st phase is the data collection phase, the massive input data need to be somehow stored and made conveniently available for later processing, this is very similar to the ADAS application's data import phase. Most of the data during this phase will be written once, read many times, and the operations are mostly sequential I/O. Low-cost SDS with great scalability is very suitable for this phase. AI applications will consume data from this pool of data.

The 2nd phase is raw data processing, in this phase raw data will go through preliminary processing to get ready for analytic applications and tend to be stored in a more centralized place such as cloud or on-prem primary storage. The data will later be accessed more randomly and frequently, while the amount of data will be around

a magnitude lower than raw data but still very large. This will require both high capacity and higher performance. A tiered storage solution with a high-end enterprise primary storage combined with a low-cost capacity tier will be suitable for this purpose. Flexible access protocol is a key value for this phase, as this data lake will be shared by all kinds of different applications.

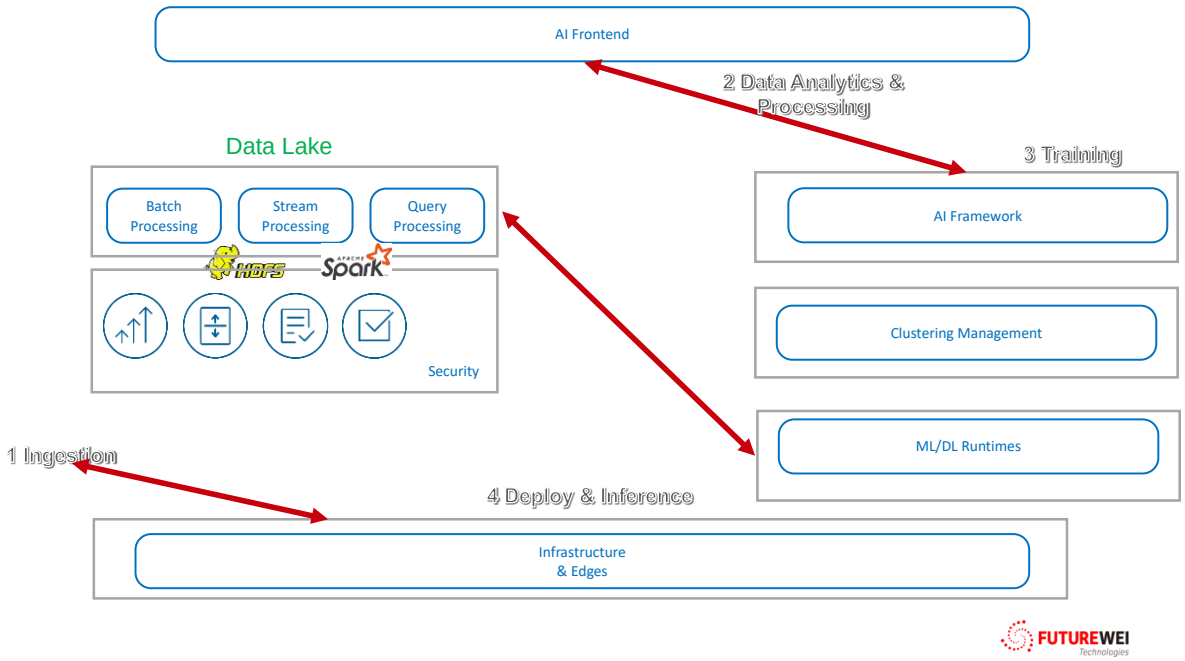
The 3rd phase is where analytic applications do their job, huge AI computing power is required, as well as high-performance local storage for intermediate results. HCI cluster with SSD-backed local storage is suitable for most scenarios. During this phase, if the pre-processed data is accessible through some high-performance channel such as 100g NVMeOF, applications can directly load data from the data lake, otherwise, data need to migrate from the phase 2 storage to a high performance local or nearby high-speed storage tier.

7.3 THE ENERGY SECTOR

The offshore oil and gas industry has changed rapidly in recent years, with new technologies being adopted by the energy sector to meet the challenges of a digital economic landscape. Artificial intelligence is an exciting new technology field and can be applied to the oil & gas industry to save substantially by simulating real project scenarios.
There are several phases for oil & gas company to explore the capabilities of AI:

1. Data collection and ingestion phase
2. Data analytics and processing
3. Training
4. Central and distributed model deployment and inference

To harvest the capabilities of AI/ML, a large amount of data is needed for training purposes. It is also desirable to build a big data lake for analytics and AI/ML. Phase 1 generally is for this purpose. Once data have been ingested by the data lake, a data analytics and processing data platform shall be provided to preprocess data and clean data for machine learning training purposes. Phase 3 is the actual training phase. At phase 4, either a centralized data model is deployed at a data center or a distributed model is deployed at the edge for inference purposes.



8 CONCLUSION

In this document, we first introduce the AI ecosystem and market trend, then introduce AI applications in many sectors and industries. Like other IT systems, AI infrastructure can be built based on on-prem clouds or public clouds. We observed that many startup enterprises tend to use the SaaS capabilities provided by public clouds, and many larger enterprises are using on-prem infrastructures.

Through these applications, we summarize the types of workloads. Although there is not a definitive workload benchmark for AI workloads, we listed several relevant benchmarks that can help the readers to benchmark an AI infrastructure. The benchmark selection will still be a subject being debated in the industry.

A general AI infrastructure is proposed. The architecture is open and based on open-source software. Based on each user's needs, the architecture can be evolved from low-end, mid-range, to high-end configurations.

Finally, we used several popular applications to show that the reference architecture is valid.

9 REFERENCES

[1] The Linux Foundation Projects, "Horovod.ai," [Online]. Available: https://horovod.ai/.

[2] M. Turck, "Resilience and Vibrancy: The 2020 Data & AI Landscape," 30 Sept 2020. [Online]. Available: https://mattturck.com/data2020/.

[3] Grand View Research, "Customer Experience Management Market Size $23.6 Billion By 2027," Feb 2020. [Online]. Available: https://www.grandviewresearch.com/press-release/global-customer-experience-management-cem-market.

[4] [Online]. Available: https://www.law.com/legaltechnews/2020/08/12/brave-new-world-how-ai-tools-are-used-in-the-legal-sector/?slreturn=20210027160243.

[5] [Online]. Available: https://www.ibm.com/downloads/cas/NAJXEKE6.

[6] [Online]. Available: https://hbr.org/2020/11/ai-can-make-bank-loans-more-fair.

[7] [Online]. Available: https://www.idc.com/getdoc.jsp?containerId=EUR246532420.

[8] [Online]. Available: https://www.uipath.com/rpa/intelligent-process-automation.

[9] K. Lyons, "YouTube brings in $5 billion in ad revenue as Alphabet and Google bounce back," 29 10 2020. [Online]. Available: https://www.theverge.com/2020/10/29/21531711/google-alphabet-ad-revenue-youtube-waymo-cloud-search.

[10] Grand View Research, "Smart Education And Learning Market Worth $680.1 Billion By 2027," Apr 2020. [Online]. Available: https://www.grandviewresearch.com/press-release/global-smart-education-learning-market#:~:text=Smart%20Education%20And%20Learning%20Market%20Worth%20%24680.1%20Billion%20By%202027,-April%202020%20%7C%20Report&text=The%20global%20smart%20education%20and,17.

[11] Grand View Research, "Smart Education And Learning Market Size, Share & Trends Analysis Report By Age, By Component (Hardware, Software, Service), By Learning Mode, By End User, By Region, And Segment Forecasts, 2020 - 2027," Apr 2020. [Online]. Available: https://www.grandviewresearch.com/industry-analysis/smart-education-learning-market.

[12] [Online]. Available: https://www.iflexion.com/blog/artificial-intelligence-real-estate.

[13] [Online]. Available: https://ash.harvard.edu/files/ash/files/artificial_intelligence_for_citizen_services.pdf.

[14] CIODIVE, [Online]. Available: https://www.ciodive.com/news/capital-one-breach-raises-questions-about-security-and-cloud-first-strategi/560129/.

[15] [Online]. Available: https://emerj.com/ai-sector-overviews/artificial-intelligence-in-insurance-trends/.

[16] [Online]. Available: https://www.bankrate.com/insurance/artificial-intelligence-meets-the-insurance-industry/.

[17] [Online]. Available: https://www.forbes.com/sites/robtoews/2020/08/26/ai-will-revolutionize-healthcare-the-transformation-has-already-begun/?sh=2f0c85c9722f.

[18] [Online]. Available: https://www.genengnews.com/magazine/314/ai-in-the-life-sciences-six-applications/.

[19] Grand View Research, "Intelligent Transportation System Market Size, Share & Trends By Type (ATIS, ATMS, ATPS, APTS, EMS), By Application (Traffic Management, Public Transport), By Region, And Segment Forecasts, 2020 - 2027," 2020. [Online]. Available: https://www.grandviewresearch.com/industry-analysis/intelligent-transportation-systems-industry#:~:text=Report%20Overview,5.8%25%20from%202020%20to%202027..

[20] Plunkett Research, "Transportation Industry Statistics and Market Size Overview, Business and Industry Statistics," 2019. [Online]. Available: https://www.plunkettresearch.com/statistics/Industry-Statistics-Transportation-Industry-Statistics-and-Market-Size-Overview/.

[21] Y. D. a. S. S. M. Piero Molino, "Introducing Ludwig, a Code-Free Deep Learning Toolbox," 11 Feb 2019. [Online]. Available: https://eng.uber.com/introducing-ludwig/.

[22] V. Panyam, "Introducing Neuropod, Uber ATG's Open Source Deep Learning Inference Engine," 8 June 2020. [Online]. Available: https://eng.uber.com/introducing-neuropod/.

[23] A. S. a. M. D. Balso, "Meet Horovod: Uber's Open Source Distributed Deep Learning Framework for TensorFlow," Uber, 17 10 2017. [Online]. Available: https://eng.uber.com/horovod/.

[24] A. Gale, "Introducing Flyte: A Cloud Native Machine Learning and Data Processing Platform," 7 Jan 2020. [Online]. Available: https://eng.lyft.com/introducing-flyte-cloud-native-machine-learning-and-data-processing-platform-fb2bb3046a59.

[25] C. Donnelly, "Uber backs hybrid cloud as route to business and geographical expansion," 7 11 2018. [Online]. Available: https://www.computerweekly.com/news/252452059/Uber-backs-hybrid-cloud-as-route-to-business-and-geographical-

expansion#:~:text=To%20achieve%20this%20scale%2C%20Uber,storage%2C%20database%20and%20GPU%20workloads..

[26] [Online]. Available: https://www.forbes.com/sites/cognitiveworld/2019/07/05/how-ai-is-transforming-agriculture/?sh=616bd49c4ad1.

[27] "SPC-1 Benchmark," [Online]. Available: http://spcresults.org/benchmark-results-spc1.

[28] "IO500 Benchmark," [Online]. Available: https://www.vi4io.org/io500/start?fields=information__system,information__institution,information__storage_vendor,information__filesystem_type,information__client_nodes,information__client_total_procs,io500__score,io500__bw,io500__md,information__data,inf.

[29] "FIO documentation," [Online]. Available: https://fio.readthedocs.io/en/latest/.

[30] "FIO project source," [Online]. Available: https://github.com/axboe/fio.

[31] "FIO script examples," [Online]. Available: https://github.com/axboe/fio/tree/master/examples.

[32] "GPU Direct Storage Guide," [Online]. Available: https://docs.nvidia.com/gpudirect-storage/troubleshooting-guide/index.html.

[33] "Nvidia GPU Wiki," [Online]. Available: https://en.wikipedia.org/wiki/List_of_Nvidia_graphics_processing_units.

[34] "Tesla GPU Wiki," [Online]. Available: https://en.wikipedia.org/wiki/List_of_Nvidia_graphics_processing_units#Tesla.

[35] "AI Benchmark," [Online]. Available: https://pypi.org/project/ai-benchmark/.

[36] [Online]. Available: https://www.forbes.com/sites/robtoews/2020/08/26/ai-will-revolutionize-healthcare-the-transformation-has-already-begun/?sh=2f0c85c9722f.

[37] D. H. Lei. [Online]. Available: https://drive.google.com/file/d/12JY_mIdBMFzaqP-lppHkUdjgWxMVRBXu/view.

[38] [Online]. Available: https://arxiv.org/pdf/1910.05930.pdf.