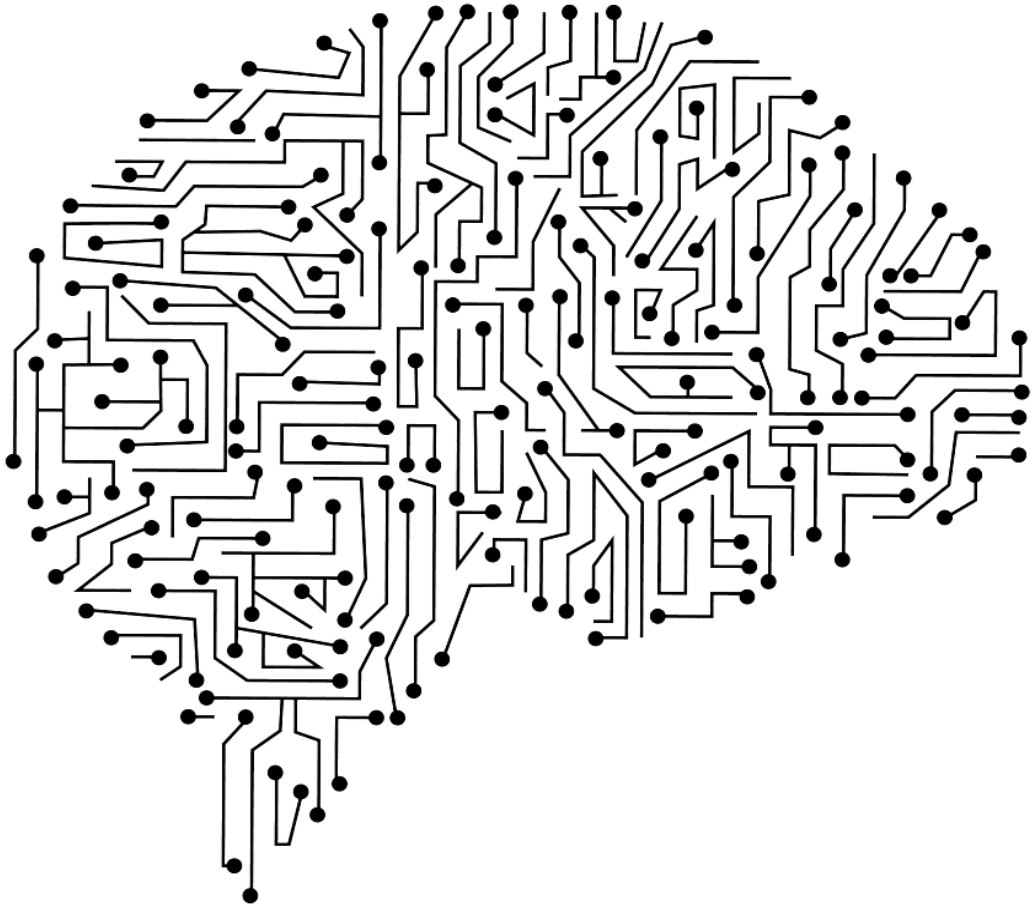
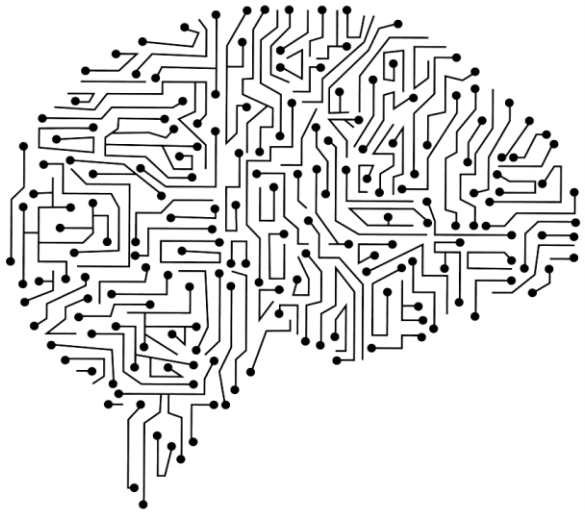


Enterprise AI and its supporting technical infrastructure

What is changing and how we can help our clients from an IT infrastructure point of view

PAULA QUEIROZ | September 2024



About this presentation

This presentation aims to provide an overview of what AI, specifically enterprise AI is and how businesses providing IT infrastructure consultancy services can shape their consultants skill sets in order to provide their clients with the help required to built fully fit for purpose AI-supporting infrastructures.

In order to do this, the following path is proposed:

- What is AI
- How AI is changing enterprises
- The AI stack
- Responsible AI
- IT infrastructures and AI
- IT infrastructure skill sets required

AI… what is what?

Defining Enterprise AI

AI: the dictionary definition

The Cambridge Dictionary defines AI (Artificial Intelligence) as:

The use or study of computer systems or machines that have some of the qualities that the human brain has, such as the ability to interpret and produce language in a way that seems human, recognize or create images, solve problems, and learn from data supplied to them.

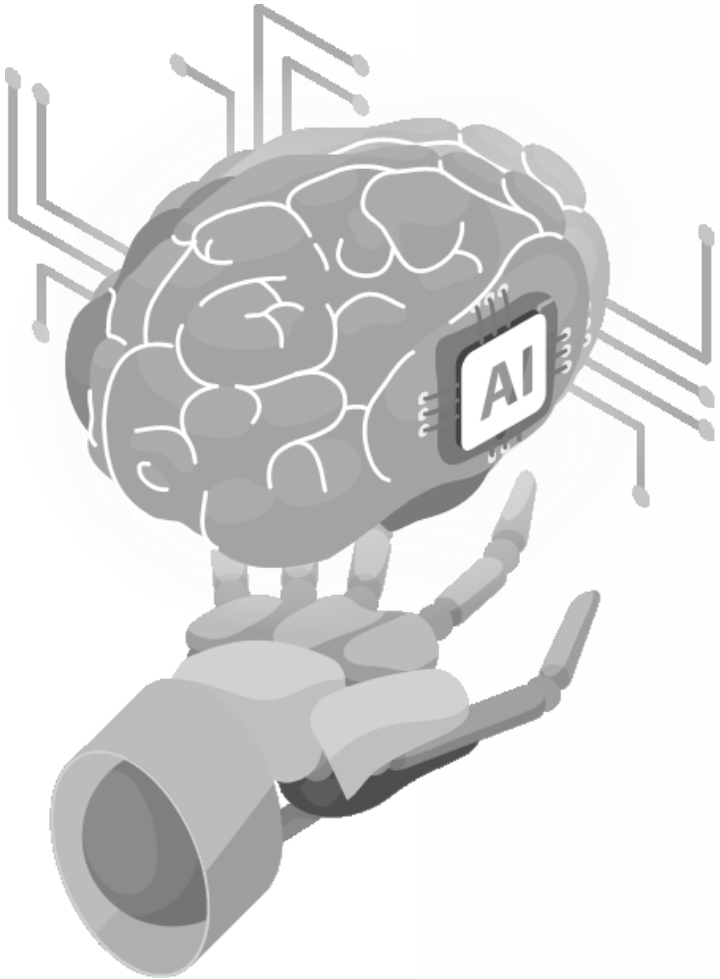
Consumer and Enterprise AI

Enterprise AI is a category of enterprise software integrating AI-enabled technologies and techniques for the enhancement of business functions. It differs quite a bit from consumer AI, which has the main (very high level) characteristics:

- Consumer AI focuses on user experience.
- It interacts directly with the end-consumer in a user-friendly way.
- This interaction is done through messengers, emails, etc, giving the user an experience similar to human interaction.
- Consumer AI aims to make highly “behind the curtains” functionality appear simple, pleasant and highly interactive to end-users.

Enterprise-AI, on the other hand, is organization-centric. The focus is on business processes, rather than on people.

Defining Enterprise AI



Consumer and Enterprise AI (continued)

Enterprise AI goes far beyond simple automation of tasks and provides the means to solve complex problems requiring human-like cognition.

This human-like “cognition” allows enterprise AI to go beyond problem solving and truly support digital transformation through its ability to understand large sets of enterprise data and processes, as well as human behaviours, allowing it to be proactive rather than just reactive.

Given the inherent complexity of enterprise AI, it requires very robust backends that provide enough compute to execute highly complex algorithms over very large amounts of data, spread across different storage, following the principles of classical information security as well as the new concept of “responsible AI”.

On this last paragraph, whilst it may be tempting to think as enterprise AI as “moving away from infrastructure services”, that is not the case.

AI relies on an infrastructure which, to keep up with how it has been evolving, will have to be far more robust and fine-tuned than the pre-enterprise AI infrastructures we are all familiar with.

How AI is changing the Enterprise

Benefits of AI in the Enterprise

The following provides a non-exhaustive, high-level overview of how AI can help enterprises in a number of sectors.

Again, the following is just to give us a brief idea, AI can provide benefits to enterprises in virtually all sectors, and those benefits go far beyond those listed.

AI and the enterprise, by sector

Healthcare

- AI-driven diagnostic tools can analyze medical images more quickly and accurately than human radiologists, reducing diagnostic errors and improving patient outcomes.

Finance

- AI algorithms can automate risk assessment and fraud detection, allowing real-time decision-making that protects assets and reduces financial losses.

Energy

- AI can predict energy demand peaks and optimize grid distribution, enhancing power generation and distribution efficiency.

Telecommunications

- AI can enhance network optimization and fault detection, ensuring better service quality and operational efficiency.

Cybersecurity

- AI systems can identify and react to security threats faster than traditional software, improving data protection.

Legal

- AI can automate document analysis and research, reducing the time and cost of legal operations.

How enterprises are leveraging the benefits of AI

The following provides two known examples of how two giants we are all familiar with are already leveraging the power of AI.

Again, the following is just to give us a brief idea of how some enterprises are using AI to improve business.

1. PayPal and fraud detection

Overview: PayPal employs AI to enhance its fraud detection capabilities and ensure secure user transactions. The AI system analyzes transaction data in real-time to identify and prevent fraudulent activities.

Example: The AI system flags a fraudulent transaction attempt using a stolen credit card due to unusual spending patterns and location. The transaction is blocked, and the user is notified immediately.

Results:

- **Fraud Prevention:** PayPal has significantly reduced the incidence of fraud, protecting customers and merchants.
- **Efficiency:** Automated fraud detection allows quick response times, minimizing financial losses.
- **User Trust:** Enhanced security measures have increased user trust and confidence in using PayPal for online transactions.

2. Amazon and the supply chain management

Overview: Amazon utilizes AI to optimize its supply chain management, ensuring timely delivery of millions of products worldwide. The company employs AI for demand forecasting, inventory management, and warehouse automation.

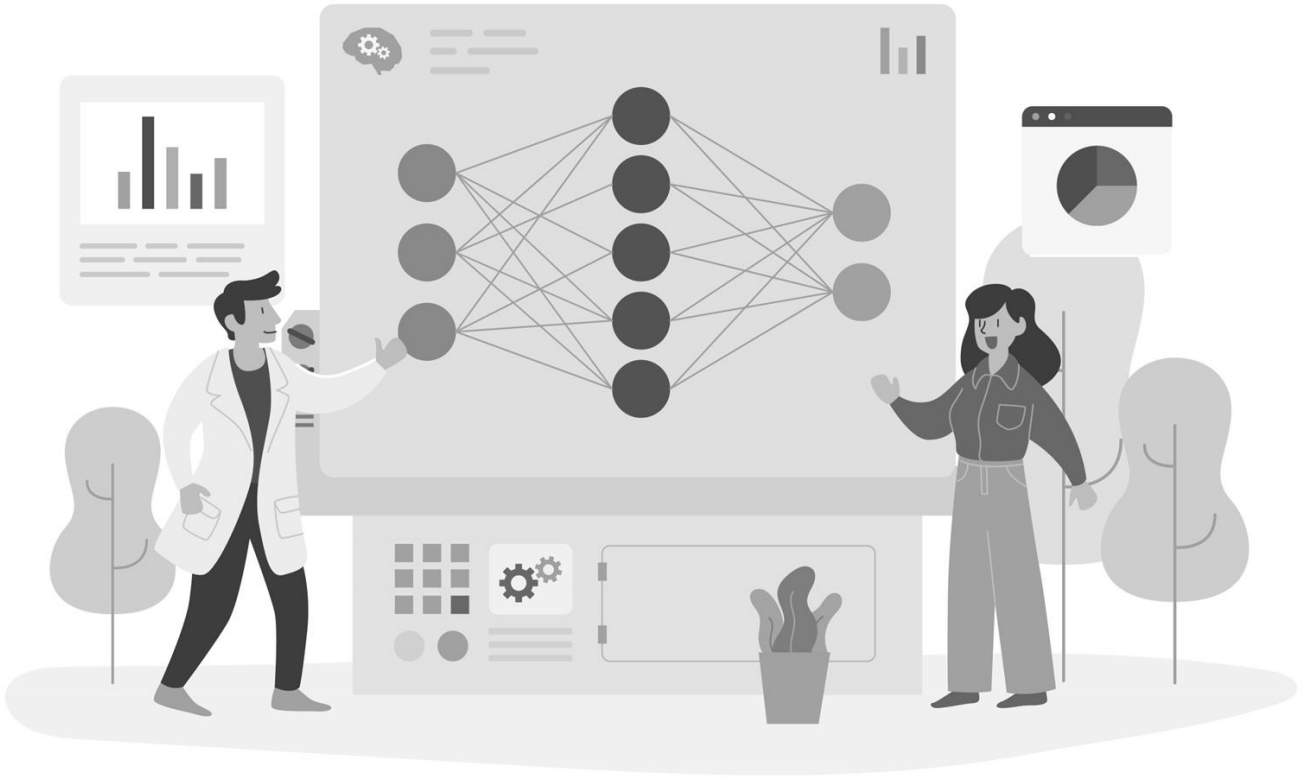
Example: When a customer places an order for a popular product, AI algorithms predict the demand and ensure the product is available in the nearest warehouse, allowing for same-day or next-day delivery.

Results:

Efficiency: AI-driven automation has reduced order processing time and minimized human error.

Cost Savings: Optimized inventory management reduces holding costs and minimizes stockouts and overstock situations.

Customer Satisfaction: Improved accuracy in demand forecasting and quicker order fulfillment have led to higher customer satisfaction and loyalty.



The AI Stack

The AI Stack

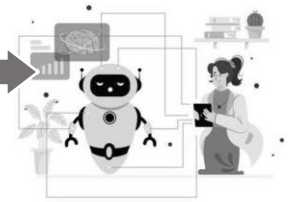
Here is an hierarchically illustrated, brief description of the meaning of:

- **Artificial Intelligence**
- **Machine Learning**
- **Neural Networks**
- **Deep Learning**
- **Generative AI**

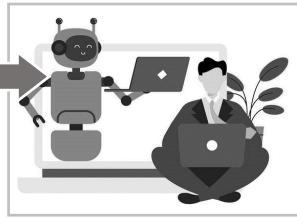
Artificial Intelligence (AI): a broad term encompassing the development and use of computer systems capable of performing tasks that typically require human intelligence.

Machine Learning: is a subset of AI that trains systems to learn from data and make decisions or predictions based on patterns.

Neural Networks: a type of machine learning algorithm that mimics the structure and function of the human brain, allowing AI systems to learn and process complex data.



Deep Learning: a subfield of machine learning that uses neural networks with multiple layers to learn and extract features from data.



Generative AI: a subset of AI that focuses on generating new content, such as text or images, based on patterns learned from data.

The AI Stack

In order to understand the basic infrastructure architecture and engineering skill sets required to design and maintain the infrastructure an enterprise AI deployment will rely on, let's focus on the basic components of the AI stack.

Run:ai gives us the following as the most basic five components of the stack:

- **Data storage and management**
- **Compute resources**
- **Data processing frameworks**
- **Machine Learning frameworks**
- **MLOps platforms**

The description of each can be found in the appendix, but, as we are focusing on the bottom of the stack, that is, infrastructure here, we will only be looking into data storage, compute resources and MLOps platforms here.

The AI Stack (from an IT infrastructure point of view)

1. Data Storage and Management

Talking about AI is talking about data, a lot of data, the more data the better.

AI relies on copious amounts of data in order to learn and validate knowledge, which makes data storage and management absolutely vital for any AI solution.

Whist we any not going to go into much detail (more information concerning « data » can be found in appendix), here is something to be kept in mind:

- Data storage will potentially involve databases, data warehouses, and/or data lakes.
- I might be on-premises or cloud-based.
- Proper data management includes ensuring data privacy and security, following the data management lifecycle (see appendix).
- Data will very likely be in various formats and from various sources.

2. Compute Resources

AI is compute intensive and, whether on-premises of leveraging cloud offers, may require specialized hardware such as GPUs or TPUs, briefly described as follows:

2.1. GPU computing

Consists of using graphics processing units for tasks beyond traditional graphics rendering.

This is effective due to the GPU's capability to perform parallel processing.

A GPU consists of thousands of smaller cores that work in parallel.

This architecture makes GPUs well-suited for tasks that:

- Involve large data sets that require extensive processing.
- Are dividable into smaller units of work the GPU can execute concurrently.
- Are highly repetitive.

2.2. TPU Computing

A Tensor Processing Unit (TPU) is specialized hardware that significantly accelerates machine learning (ML) workloads.

It is meant to handle compute-intensive operations of deep learning algorithms, and provides a more efficient and faster way to execute large-scale ML models.

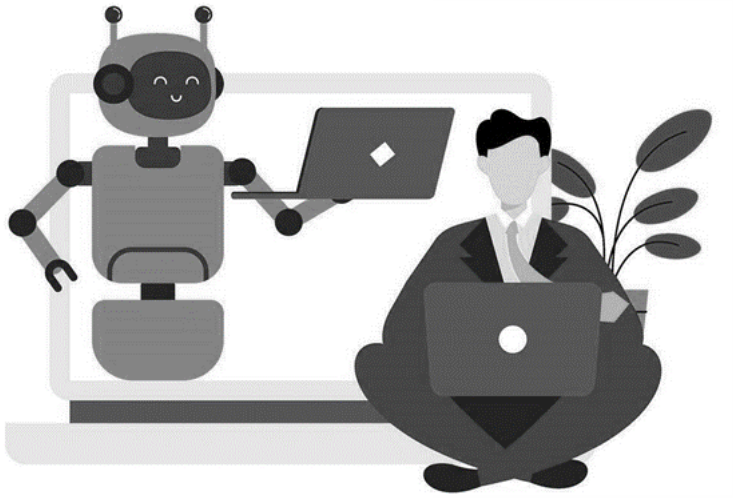
3. MLOps Platforms

MLOps involves the principles and practices of automating and streamlining the machine learning lifecycle, from data collection and model training to deployment and monitoring.

MLOps platforms help manage this lifecycle, including version control for models, automated training and deployment pipelines, model performance tracking, and facilitating collaboration between different roles (data scientists, ML engineers, operations, etc).

Responsible AI

Responsible AI



Responsible AI is an approach to AI based on ethical and legal grounds which aims to ensure that AI is safe and trustworthy, putting emphasis on transparency and tackling the threat of AI bias.

The ethics of AI are a big challenge we currently face but which are becoming increasingly important with the expansion of AI and must be integrated in the development and use of AI technology.

As proposed by many vendors offering AI solutions:

- **Ethical AI is based around societal values and trying to do the right thing.**
- **Responsible AI is the tactical approach to ethical AI, relating to the way and use technology and tools are developed and used in an ethical manner.**

Confronting ethical concerns means engaging with their ramifications with foresight and commitment. Embedding responsible AI principles is essential to AI evolution in a direction that benefits all.

There is no a fixed, universally agreed-upon set of principles for AI ethics, but several guidelines have recently started to emerge.

The next section of this presentation will explore some of the most common principles we are now expected to have in mind in the development and use of AI.

Responsible AI

The following principles for responsible AI are those proposed by the International Organisation for Standardisation (ISO):

1. Fairness

Datasets used for training the AI system must be given careful consideration to avoid discrimination.

2. Transparency

AI systems should be designed in a way that allows users to understand how the algorithms work.

3. Non-maleficence

AI systems should avoid harming individuals, society or the environment.

4. Accountability

Developers, organizations and policymakers must ensure AI is developed and used responsibly.

4. Privacy

AI must protect people's personal data, which involves developing mechanisms for individuals to control how their data is collected and used.

6. Robustness

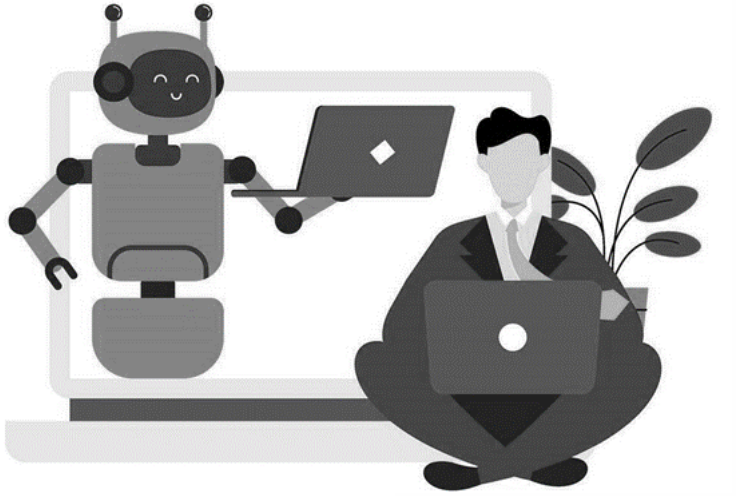
AI systems should be secure – that is, resilient to errors, adversarial attacks and unexpected inputs.

7. Inclusiveness

Engaging with diverse perspectives helps identify potential ethical concerns of AI and ensures a collective effort to address them.

AI and IT infrastructure technical consultant profiles

Enterprise AI skillsets



In this section of the presentation we will be discussing what the desired skillsets for infrastructure technical consultants working in AI-related projects should look like.

Different enterprises may request additional skills, or even less skills than those presented, so this is just a broad overview of what these consultants are expected to include in their CVs.

As we have seen, the following as core requirements for any AI implementation, so they should be included:

- **Storage**
- **Compute**
- **Information-security** (including legal and regulatory compliance)

Note: a large enterprise will probably have the above split into different roles (ie: storage engineer, information security consultant, etc).

Enterprise AI and Technical Infrastructure Profiles

Data Engineer

Data engineering is the practice of designing and building systems for collecting, storing, and analyzing data at scale. It is a broad field with applications in just about every industry.

Organizations have the ability to collect massive amounts of data, and they need the right people and technology to ensure it is in a highly usable state by the time it reaches data scientists and analysts.

Some of the common tasks a data engineer might perform when working with data include:

- Acquire datasets that align with business needs
- Develop algorithms to transform data into useful, actionable information
- Build, test, and maintain database pipeline architectures
- Collaborate with management to understand company objectives
- Create new data validation methods and data analysis tools
- Ensure compliance with data governance and security policies

A data engineer's skills will include the majority of the following:

- Coding, automation and scriting (ie SQL, NoSQL, Python, Java, etc)
- Relational and non-relational databases
- ETL (extract, transform, and load) systems (ie Xplenty, Stitch, Alooma, etc)
- Data storage
- Automation and scripting
- Machine learning
- Big data tools (ie Hadoop, MongoDB, Kafka, etc)
- Cloud computing (ie AWS, Google Cloud, etc)
- Data security

Enterprise AI and Technical Infrastructure Profiles

Storage Engineer

A storage engineer is responsible for designing, implementing, and deploying shared system resources, like disks and networking. Storage engineers also work on performance analysis and troubleshooting, data recovery strategies for critical systems, and data backup measures.

Some of a storage engineer's responsibilities will include:

- Utilising in-depth experience in managing critical backup infrastructure and processes
- Gathering requirements, installation, configuration, and testing of backups in a high availability enterprise system
- Provide daily support of the storage environment and backups
- Work on storage related tasks such as volume creation, LUN provisioning and data migration
- Provision storage and monitors capacity
- Perform all maintenance, configuration and deployments to the SAN system

A storage engineer will have most of the following skills:

- Responsible for the day-to-day administration of the storage area network (SAN)
- Provision storage and monitors capacity
- Perform all maintenance, configuration and deployments to the SAN system
- Work on storage related tasks such as volume creation, LUN provisioning and data migration
- Provide daily support of the storage environment and backups

Enterprise AI and Technical Infrastructure Profiles

Datacentre Engineer

A datacentre engineer acts as a manager for both the hardware and software that an information datacentre utilises. This highly technical position requires a variety of information technology and IT-oriented competencies.

A datacentre engineer will be expected to perform must of the following tasks:

- Instals, manages and troubleshoots the information systems on which the datecentre runs
- Provides expertise on the correct methods to optimally run the datacentre
- Acts as a company-facing information technology expert as well, internally diagnosing and solving technical problems to ensure the data centreruns smoothly
- Gnerally also responsible for selecting and installing the hardware that a datacentre uses

More straight to the point, the datacentre engineer’s main responsibilities are:

- Hardware and software management
- Securing the network's information
- Troubleshooting and maintenance
- Optimising datacentre operations

Enterprise AI and Technical Infrastructure Profiles

High Performace Compute (HPC) Systems Engineer

The HPC Systems Engineer role has the overall responsibility to work within a team to provide a performant, reliable, and secure high-performance computing (HPC) environment. The HPC Systems Engineer will be involved in various aspects of designing and engineering our HPC system as well as be responsible for managing day-to-day operations and maintenance activities.

Their primary job functions are:

- Establish strategies for overall support of the system
- Evaluate new hardware and software and understand potential benefits/impacts it can have in the environment
- Perform hardware maintenance
- Perform software installations and upgrades, inclusive of operating system
- Monitor overall system performance and health
- Provide support for the management of data in the environment
- Work with users to resolve problems and ensure they are able to effectively utilize the system
- Interact with both business customers and technical teams that are globally distributed and within varied time zones
- Engaging with vendors for problem resolution of existing infrastructure and discussion of roadmaps and new technologies for evaluations
- Foster a supportive work environment and maintains open, productive interactions among team and across organizations
- Build and maintain cross-organizational contacts to facilitate execution of work

Enterprise AI and Technical Infrastructure Profiles

IT Risk Analyst

IT risk analyst provides advisory services related to internal controls, risk assessments, risk management, IT controls, related standards (HIPAA, HITECH, NIST, etc.) and corrective action plans.

Some of an IT Risk Analyst's duties will include:

- Analyse and produce functional specifications arising from the requirements of users
- Develop and execute test plans for integration testing to ensure business requirements have been met
- Analyse and respond to complex queries and issues from risk managers and other users relating not only to current projects but also to runtime system functionality and future concepts/ideas
- Manages financial operations for Risk IT which includes invoice management, SOW management, month end close, inter-region journal entries, cost center investigations, professional service fee accruals, and the overall integrity of portfolio financials
- Assists in forecasting and planning cycles, including monthly/quarterly project forecasts, annual IT planning, and quarterly finance cost center forecasts
- Assists in the creation and support of the Month End Closure cycle covering IT Risk, Project management, financials and resource reporting
- Supports and participates in the evaluation of 3rd party vendor data, internal project data and internal expense management reporting
- Provides group analytical services supporting the Risk IT team

Enterprise AI and Technical Infrastructure Profiles

Cybersecurity Analyst

A cybersecurity analyst protects company hardware, software, and networks from cybercriminals.

The analyst's primary role is to understand company IT infrastructure in detail, to monitor it at all times, and to evaluate threats that could potentially breach the network.

The cybersecurity analyst continuously looks for ways to enhance company network security and protect its sensitive information.

The cybersecurity analyst is also responsible for:

- Configuring tools:

  This may come in the form of virus software, password protectors, and vulnerability management software. They will evaluate what the company needs and use these tools to protect its information.

- Reporting:

  The analyst will detail what is currently going on in the network and evaluate its strengths. One of the skills needed is learning to read these reports. They will show what is well-protected and indicate if there is any unusual activity in the network.

- Evaluate weaknesses:

  No network is fully secure, but the goal is to make it as secure as possible. Part of the job is to continuously test all company networks and find weaknesses before bad actors or external threats can compromise them.

Enterprise AI and Technical Infrastructure Profiles

AI Architect

AI architects lead the creation of a company's AI architecture with varying frameworks and deployment models to devise and execute AI architecture strategy.

AI architects work closely with other teams, including data scientists, machine learning operations, and company leadership and stakeholders.

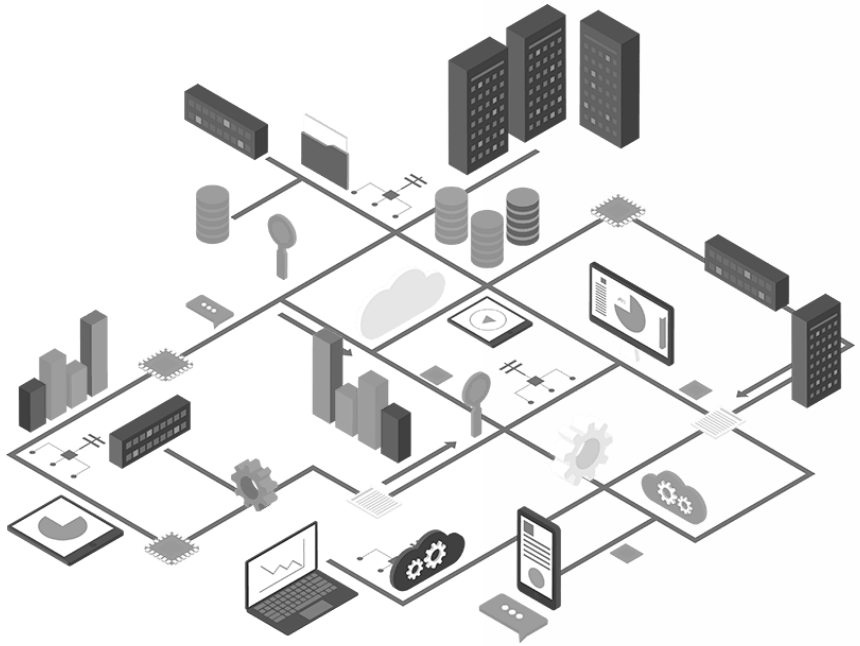
They are deeply involved in helping organizations move forward with integrating AI into their existing systems and prepare their systems for new programs and applications to keep pace with emerging trends.

Some of an AI Architect’s duties will be:

- Developing AI models, systems, and infrastructure to help drive organizational improvements and consumer products.
- Working with other IT team members, including data scientists and leaders, to support digital transformation.
- Building systems that teams, departments, or companies can integrate into existing systems.
- Developing new AI-related applications and managing programmers.
- Implementing machine learning models and converting them into application programming interfaces (APIs) for various uses.
- Helping define AI architecture and guide leaders and decision-makers in choosing compatible technologies.
- Collaborating with security professionals to manage potential risks and implement AI technologies, applications, and infrastructure in keeping with ethical policies

Enterprise AI and Technical Infrastructure Profiles

Infrastructure Architect



Infrastructure architects plan, design, assemble, and oversee the systems that run an organization's technological infrastructure.

They assess the operation's existing resources, determine any changes that may be needed, and make recommendations.

They also provide ongoing support and advice to ensure the infrastructure continues to support fluctuating demands.

Working in collaboration with IT, operations, and other teams, they develop strategies and action plans for implementing upgrades and integrations.

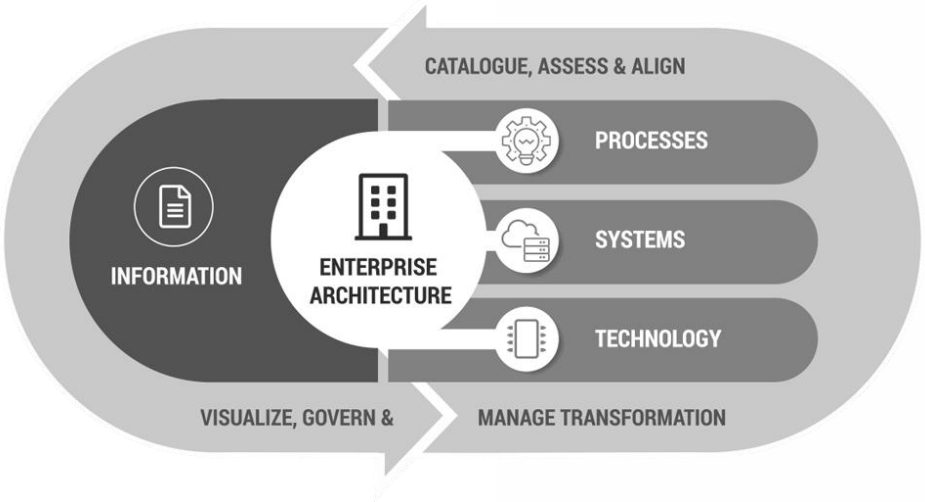
They also outline evaluation benchmarks for future assessments.

The Infrastructure Architect will have most, if not all, of the following skills:

- Strong leader and communicator.
- Good problem solving and analytical skills.
- Experience with web technologies and building enterprise architecture.
- Strong hold on designing, integrating and managing complex infrastructure solutions.
- Ability to assist in planning and support processes and adhering to best practices.

Enterprise AI and Technical Infrastructure Profiles

Enterprise Architect



The Enterprise Architect combines many different skills and experiences to address and solve business, information, and technology-related opportunities and problems.

In many activities, the Enterprise Architecture developed by the Enterprise Architect is the “glue” that integrates the project and program strategies across multiple programs and projects and ensures alignment with business strategies and drivers, and management priorities.

Enterprise Architect sets the direction and establishes the approach for integrating information applications and programs.

The Enterprise Architect’s key responsibilities include:

- Definition, implementation, and execution of the processes for the definition, maintenance, and conformance management of the Enterprise Architecture.
- Update and maintenance of the key Enterprise Architecture deliverables.
- Establishment and maintenance of contacts within business units and information system programs to understand business activities and business drivers, business requirements, solutions strategies, alternatives, etc., being considered and/or implemented.
- Architectural leadership in the resolutions of inter-program and inter-project issues.
- On-going publicity and communication of the Enterprise Architecture both within the information community and the business units.
- Ongoing research and assessment of new analysis approach for potential use within the Enterprise.

Resources

https://www.run.ai/guides/machine-learning-engineering/ai-infrastructure

https://redresscompliance.com/artificial-intelligence-an-introduction-to-ai-fundamentals/

https://www.teradata.com/insights/ai-and-machine-learning/6-machine-learning-tools-for-enterprises

https://c3.ai/what-is-enterprise-ai/

https://www.getguru.com/reference/enterprise-ai

ISO - Building a responsible AI: How to manage the AI ethics debate