Executive Insight paper

Commissioned by

Laying the foundations for Enterprise AI

An architectural approach to deep learning platforms

Freeform Dynamics, 2018

Executive Summary

About this Document

The insights presented in this document are derived from ongoing independent research, coupled with specific briefings from Fujitsu on their integrated HPC and Big Data stacks, delivered under the PRIMEFLEX brand. While specific technologies are used to illustrate how key generic principles translate to practical reality, nothing in this paper should be taken as a validation or endorsement of any product or supplier.

Technologies from the Big Data and HPC worlds are enabling the widespread adoption of Deep Learning services.

The likes of Facebook, Google, Apple, Amazon and Microsoft demonstrate the power of Artificial Intelligence (AI) on a daily basis. Indeed, how investors see the value of such companies is influenced heavily by their AI prowess. This AI imperative is now spreading to the business mainstream and will become fundamental to the operation of most industries.

But how does this translate to actionable practicalities for enterprise organizations? And what needs to be done to lay the right technology foundations to meet both immediate and future needs?

Deep Learning is key to any enterprise AI strategy

One of many technologies within the broader AI field, Deep Learning (DL) is already having significant impact in the enterprise AI arena. Exploiting developments in neural networks, DL allows organizations to build powerful inference engines that can take both decision-making and automation to the next level.

The opportunity is broad, and spans many industries

Areas of application for DL range from image/object recognition, through intelligent real-time event processing, to advanced modelling and simulation. These map onto specific solutions in many different industries.

Application demands determine the architecture required

The initial stages of implementation focus on feeding the DL system data it can learn from. The data volumes involved are often very large, and a lot of computing power is required to drive the learning process. Fortunately, technologies and approaches developed in the Big Data and High-Performance Computing (HPC) worlds can be brought together to meet the need.

Getting the converged architecture right is critical

A lot goes into designing an enterprise-class DL platform. Beyond selecting, sizing, integrating and configuring the necessary components, you need the right development environment and operational tools to deal with system, data and security management. However, reference architectures exist that can act as a blueprint for you or your supplier to build the platform to meet your needs.

AI: From popular fiction to business advantage

Today, hardly a news cycle goes by without some kind of big AI story appearing in the press. The challenge is understanding what AI means in a business context.

Until just a few years ago, most people would have typically associated Artificial Intelligence (AI) with academia and advanced research institutions, and of course with popular fiction with its androids and sentient computers.

All that has changed. Today, hardly a news cycle goes by without some kind of big AI story appearing in the press. Sure, a lot of what's reported is exaggerated and sensationalized, but underpinning the drama and the hype are serious developments that are making AI technology genuinely accessible to the enterprise mainstream.

The challenge, though, is not just distinguishing reality from fantasy, but also making sense of everything being discussed in this space. At one end of the spectrum, the term 'AI' is used to describe the sophisticated consumer profiling platforms and techniques used by the likes of Google and Facebook. At the other extreme, it has been hijacked by tech marketing departments to rebrand old, familiar automation solutions that IT teams have been using for years. In between, there's a confusing array of technologies and application types that are all frequently labelled AI.

Our aim in this paper is to cut through the noise and look at some of the tangible and practical platform and technology options that can be used today to meet real business needs.

Machine Learning, and Deep Learning are very real, with the potential to unlock significant business advantage from the huge amounts of data now available.

Covering the space exhaustively is clearly beyond the scope of any single document, so when discussing opportunities, requirements and solutions, we have chosen to focus on a specific form of AI, namely Deep Learning. This is quite rightly receiving a lot of attention in both business and enterprise IT circles, not least because of its potential to unlock significant business advantage from the increasing amounts of data now available to most organizations.

From Machine Learning to Deep Learning

Before going any further, let's explain what we mean by Machine Learning and Deep Learning, which you'll often see us shortening to ML and DL for convenience in the remainder of this paper.

Firstly, to orientate ourselves on the general landscape and provide some overall context, it is worth pointing out that DL is a subset of ML, and ML is in turn a subset of broader AI (Figure 1).

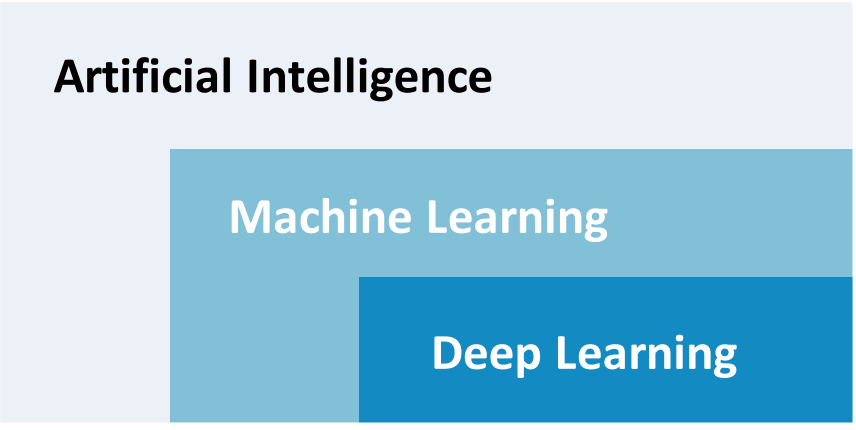

Figure 1
Putting Machine Learning and Deep Learning in context

But what do these terms mean in practice? Here is what we mean when we refer to Machine Learning:

> ***Machine Learning (ML)*** *allows software applications to take an algorithmic approach to parsing, analyzing and inferring insights from data in order to support assisted or automated identification and decision-making. In essence, an ML-based system 'learns' from training data it is given, working out how to make sense of it by example without the need to be fed explicit rules and other specifications. It can optionally continue 'learning' and refining its 'skills' as additional data is acquired.*

If you provide an ML system with enough photos of apples, it can 'learn' to recognize apples in images that it's never 'seen' before.

A simple example is image recognition. If you provide such a system with enough photos of apples, it can 'learn' what an apple looks like. It will then be able to recognize apples in subsequent images that it's never 'seen' before. An example of an ML solution with a more obvious benefit is one that can learn to identify cancerous cells or lesions in a medical context to assist with diagnoses.

We might also think of a system that continuously analyses video feeds from a manufacturing production line, helping to minimize waste, cost and lost production time.

We'll talk more about specific applications a little later, but in the meantime let's turn to what we mean by DL:

> ***Deep Learning (DL)*** *systems take a form of ML called an* ***artificial neural network*** *and extend it by adding more layers (depth) to the analysis and learning process. A key aspect of DL is that the various inputs into the learning and inference processes may themselves be based on the analysis and output from previous layers of processing.*

Sounds complicated? Well, one of the best ways to understand the principle of Deep Learning is to think of how the human brain makes sense of the world and learns from inputs and experiences. Our eyes feed images into our brain that are processed and interpreted based on patterns, timings, and knowledge of what we have seen before. It's the same with language and other sounds coming in through our ears. Then we put the multiple inputs and inferences together – e.g. if someone asks if we are hungry and lobs a green object at us, we know to catch the apple. If they curse us and hurl something that looks grey and hard, then we know to duck to avoid the rock.

To understand the principles of Deep Learning, think how the human brain makes sense of the world, learning from input and experience.

This is clearly a simplification of what happens in humans – our brains will also take into account whether we know the person doing the throwing, what we can infer about their intentions from their expression and posture, the general context in which the item is thrown, and of course history – events of the last few minutes, previous similar experiences, and so on.

The point is that if you were to try to write a set of precise, logical and prescriptive rules for how to react when someone throws something in your direction, you would find it very difficult – it's about assessing all available inputs and information and making a judgement call. That's pretty much the principle behind DL.

We'll touch on the deeper technical and practical requirements associated with the 'teaching' and 'execution' phases of DL system implementation shortly, but before this, let's take a look at the mainstream potential of such solutions.

Why is Deep Learning happening now?

Deep Learning improves with more data and greater compute capability. Advances in both these underlying technologies have therefore greatly boosted DL.

As discussed, DL adds more layers to the analytical algorithm, in a loose emulation of how the human brain learns. This increases the processing load and requires more data, so in practical terms it could not achieve success until the necessary HPC and Big Data capabilities had evolved, coming down in cost and rising in usability. Both these areas have advanced immensely in the last few years.

Part of the cost reduction was due to the use of commodity hardware, but there was also the key realization that the matrix mathematics used in DL closely resembles the maths used in generating graphics within a PC. This allows relatively inexpensive HPC systems to be built around graphical processing units (graphic chips, or GPUs).

The result has been a virtuous circle of advancements in DL hardware, software and know-how: Cheaper and more capable DL technology has enabled researchers to develop more and better DL algorithms. Improved algorithms help DL work better in the enterprise, and so on.

Scope of relevance and opportunity

To give a feel for the potential, Figure 2 shows how some fundamental DL capabilities already map onto a range of industry-specific use cases and application types.

Figure 2
Mapping of common capabilities onto industry use cases

	Manufacturing	Healthcare / Life Sciences	Financial Services
Advanced modelling & simulation	Production, logistics and maintenance planning	Drug design and treatment optimisation	Customer/market insights, risk analysis/modelling
Intelligent, real-time event processing	Production, logistics and maintenance execution	Health monitoring and incident alerting	Market tracking, fraud and security monitoring
Image/object analysis & recognition	Production monitoring and quality management	Rapid and enhanced diagnostics	Insurance risk and claims assessment
Natural language analysis/interaction	Hands-free, eyes-free setup and assistance	Assisted and self-service patient diagnostics	Market sentiment analysis, customer self-service

This picture is far from exhaustive: if we tried to map all DL capabilities across all business sectors and use cases, the table would need to be deep, wide, and probably three-dimensional. There are also generic use cases such as customer engagement, security, and process optimization that apply in almost any industry. The key is that DL technology is now far from niche, so if you haven't done so already, you should be thinking about where and how it could play a role in your plans and activities.

Business software and SaaS platforms will increasingly embed Deep Learning 'skills'.

Understand too that it is impossible to stop AI entering your world and impacting your business. If you use software applications or services such as ERP, CRM, and line of business solutions, then you will increasingly see DL-based 'skills' embedded in them. Others are also exploiting DL for advantage, including your suppliers, customers and competitors. Closer to home for IT teams, there is nowadays a DL aspect to many cybersecurity solutions and to the tools you use for monitoring and management.

In the remainder of this paper, however, we will focus on what you need to think about if you are building and/or implementing your own solutions, with a particular focus on Deep Learning and the platforms and infrastructure required to support it.

Deep Learning implementation practicalities

In order to understand the systems and platform level requirements to support Deep Learning, we first need to walk through the essentials of assembling a solution.

At its simplest, DL is typically implemented via a software framework such as Tensorflow, Caffe or Theano. On this we build a DL model based on the problem to be solved. At a system level we need to accelerate that immensely complex model, or more accurately, the software framework encoding the model.

Training and execution

In the initial stage, the focus is on training: feeding the system large amounts of data that it can learn from.

In the initial stages of implementation, the focus is heavily on training. This typically depends on feeding large amounts of data into the system, which uses a deep neural network to learn from it. The kind of data we are talking about here ranges from structured text and transaction records, through unstructured documents and free-text sources, to multi-media files, including images, audio and video.

The objective is usually to derive models that encapsulate the knowledge and insights accumulated from the learning process. These can then be deployed in the production environment to make complex inferences and decisions at execution time, based on live inputs, i.e. incoming data and events. Details of the process, activities and flows involved will vary depending on the exact nature of the environment and application, but here's a high-level view of some of the key elements (Figure 3).

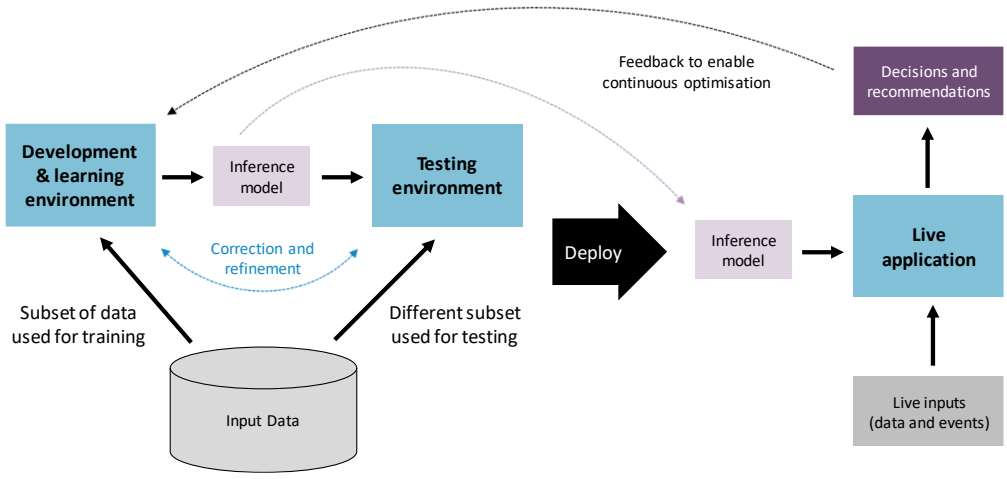

Figure 3
Stylized representation of a Deep Learning environment

The example we see here is not definitive, as numerous approaches exist for constructing, training and building applications. Sometimes the live application is physically separate, e.g. it may even run on a relatively low-spec edge device. On other occasions the line between development, test and execution is relatively blurred.

The need for computing power and data handling capacity

Whatever the environment and associated processes, training activities in particular frequently require a lot of computing power and the ability to handle very high volumes of data. As a general rule, the more sample data that's used during training, the more accurate and consistent the results produced. This in turn clearly influences the level of value and utility of an application, and ultimately the ROI it delivers.

So how does all the above translate to specific platform requirements?

A platform technology perspective

Deep Learning may sound very demanding from a technology perspective, yet has seen significant growth in recent years, both in the amount of research carried out and in its application to real-world uses. That's because, as mentioned earlier, DL has been able to leverage other great leaps that have been made in recent years, in high-performance computing (HPC) and Big Data technologies. Let's explore this further.

Great leaps have been made in recent years in both HPC and Big Data technologies.

The advantage of convergence

HPC and Big Data stacks have both been around for many years, and many large organizations will have some experience with them. HPC has traditionally been used where applications are very compute-intensive – common examples are modelling, simulation and complex rendering. Big Data solutions, on the other hand, are aligned with data-intensive applications, e.g. the mining of large customer, social media or telemetry data sets for patterns and trends to support business decision-making.

In the HPC world, we have seen key advances in the hardware and systems software domain. Here DL platforms will benefit from all of the work done to optimize servers for parallel and vector compute efficiency. Those DL frameworks can run faster on systems with more CPU cores, larger/hierarchical caches, more and faster memory, and GPUs incorporated as an inherent element of the architecture. We then have faster interconnects with lower latency to support parallel scale-up, and advances in the software stack to better enable development, tuning, execution and management.

In the world of Big Data, some of the most relevant advances are more to do with platform software. Open source software stacks such as Hadoop, from the likes of Cloudera®, Hortonworks® and MapR®, and self service analytics capability from Datameer®, have been put through their paces in an enterprise analytics context over the past few years. As their use has expanded and deepened, the ecosystems around them have expanded too. They have long provided the scalability, manageability, resilience and security needed for enterprise use, and now the tools and APIs required to both exploit and manage Big Data environments have matured considerably as well.

Building on familiar technologies and solutions means you can re-use existing knowledge and skills.

The convergence of these two hitherto largely separate worlds, and the developments taking place within them, is essentially what turns academic DL concepts into practical enterprise tools. Building on familiar technologies and solutions also has advantages from a resourcing perspective – a lot of existing knowledge and skills can be re-used.

From HPC and Big Data stacks to Deep Learning platforms

It's outside the bounds of this paper to go into the architecture of an Enterprise DL Platform in detail. Suffice it to say that reference architectures exist to build a DL environment from scratch to meet specific needs. But depending on your requirements, a DL platform can also be constructed by simply integrating standard HPC and Big Data stacks in an appropriate manner so they work together seamlessly. Either way, an enterprise platform for DL can benefit from all the historical and ongoing advances in HPC and Big Data technology.

Increasingly though, and quite apart from DL, there's been a growing need to deal with larger data-sets in an HPC context, and to perform more demanding processing

within Big Data environments. If you have encountered this in your organization, you will be familiar with the concept of 'stretching' inherent capabilities to cope with requirements outside of a platform's natural scope.

The challenge is that the more you stretch, the more you run into practical, performance and functional constraints, and the more you end up with costly work-arounds and compromises. In order to combine powerful computing with the ability to handle large data sets without compromise, we need to think a little differently. This leads us to the concept of an Enterprise DL Platform (Figure 4).

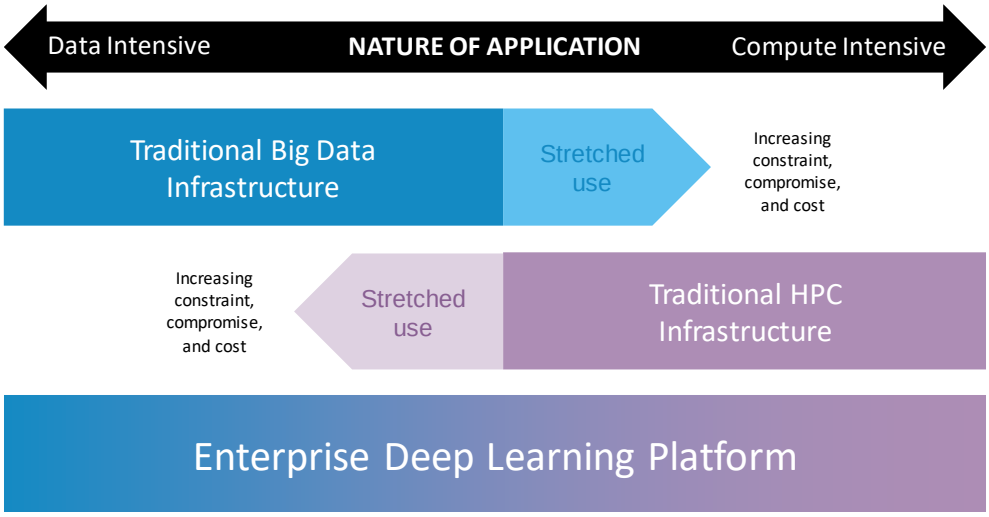

Figure 4
The need for an Enterprise Deep Learning Platform

Construct a specific platform for Deep Learning before you overstretch existing HPC and Big Data resources.

As you start to explore DL within your own organization, then if you have traditional HPC and/or Big Data environments already, the lesson here is to beware of over-stretching them. Even if they can handle the immediate requirements, one thing you can be sure of is that DL-related business demands are likely to increase over time.

It therefore makes sense to consider building a specific platform for enterprise DL. The good news is that this DL platform need not be based on a completely new set of technologies. It more represents the convergence of the HPC and Big Data worlds, and of the system-building expertise and the reference architectures that have built up around those technologies.

Cloud, core, edge

With any platform discussion nowadays, it's important to consider the cloud and hosting options. It is no different with DL. The question of where to physically deploy a DL platform or application is complex, however, not least because you may elect to split or duplicate capabilities and activities between environments, and/or change their location over the lifetime of the solution as activity and usage patterns change.

As an example, in some cases it may be preferable to run DL training locally, rather than upload huge amounts of training data from on-site systems into the cloud. In other training cases, it may be better to rent GPUs and other HPC resources in the cloud, if they will only be needed for a relatively short time.

Similarly, while some people might prefer to deploy derived models in the cloud for agility, in an IoT environment some pre-processing and filtering is often performed at the edge of the network, close to the 'things' themselves. In such cases, you may want

the DL model deployed at the edge, e.g. to enable a rapid and intelligent automated response to edge events such as sensors reporting out-of-band data.

The bottom line is that DL systems need to be as location-agnostic as possible, so you have the flexibility to mix, match and change options as needs dictate.

GDPR and data governance

Take care when you apply Deep Learning to personal data.

In the past, Big Data and HPC applications were rarely developed with compliance and regulatory issues in mind. Very few developers will have allowed for the possibility that someone might exercise their data subject rights under the GDPR, for example. Yet any organization covered by GDPR – and if you hold personal data, you are very likely covered by it – needs the ability to delete an individual's personally-identifiable data, remove it from processing, or simply not use it at all without explicit permission.

Another provision of the GDPR gives data subjects "the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her." There are exceptions to this right, but anyone applying DL to enterprise data needs to think carefully about the extent to which they are profiling identifiable individuals.

Working with suppliers and partners

In recognition of DL's emerging role, some mainstream vendors have developed fast-track approaches to system design and build to deliver comprehensive, flexible and scalable enterprise-grade platforms in this space. While we can't vouch for or recommend specific solutions, the sponsor of this paper, Fujitsu, provides a good real-world example of the kind of offerings now available from the supplier community.

Reference architectures

How you merge HPC and Big Data analytics into one system is often very use-case specific. Reference architectures are there to help.

The way in which you combine the capabilities of HPC and large-scale data analytics into a single system is often very specific to the use-case. In such scenarios it can be useful to have reference architectures or 'design blueprints' available to help put together the functional capabilities required, including the base hardware and all the required middleware.

Fujitsu has captured its experience of helping clients to assemble DL solutions and used it to create reference architectures. These can then be tuned to specific deployment scenarios such as crash simulation, climate modelling, molecular dynamics, and so on. This assists researchers and developers in these long-standing HPC use cases as they look to incorporate DL within their expanded portfolios.

> ***Example:*** *a high-tech provider of power generation and transmission solutions uses Fujitsu's integrated solution for various use cases. There are instances when the company's analysts use Fujitsu's integrated solutions to prepare and process large and often unstructured data volumes in real-time, to gain knowledge to conduct threat and risk analysis. In another instance, the use of an AI solution co-created with Fujitsu enables the automatic detection of production flaws through ML and DL capabilities.*

A key consequence of these being reference architectures for DL, rather than fixed appliances, is that although they can be delivered as packaged systems, Fujitsu's

design experts and its labs around the world can first work out the best combinations of compute (both CPU and GPU), software, storage and networking to satisfy the target workload. This means the customer does not have to design, configure, assemble and test the many elements manually, a process that can be complex, error-prone and time consuming.

Both hardware and software innovation are boosting the enterprise adoption of Deep Learning.

> ***From research into production:*** *Fujitsu's Deep Learning Unit (DLU) helps users to construct their neural networks on platforms where the bulk of the training has been done already. The experience gained from technologies such as this also helps Fujitsu construct relevant AI solutions for customers, using either their existing infrastructure as a starting point or by re-constructing solutions based on its reference architectures.*

Whether or not Fujitsu's pre-integrated HPC and Big Data stacks, delivered under the PRIMEFLEX brand, are factored into the equation depends on specific requirements. While we have warned against 'stretching' systems too far beyond their natural capabilities, it will often be the case that PRIMEFLEX for HPC or PRIMEFLEX for Hadoop will individually provide a good starting point to build out from. For a full-scope DL platform requirement, both (or at least elements of both) are likely to be utilized, but this will be covered as part of the relevant reference architecture discussion.

The delivery perspective

It's not just about the technology but also how it is delivered, supported by your chosen partner.

When it comes to solutions intended to perform effectively and efficiently over the longer term, growing and flexing along the way as requirements evolve, it's not just about the technology but also how it is delivered, supported and paid for. This is particularly true if you are driving for simplification, repeatability and standardization, and this is where suppliers with a broad set of capabilities, experience, the right mindset and financing options can come into their own.

Fujitsu, for example, brings broad and deep know-how to bear from running mission-critical infrastructure for many large enterprises. It also has a strong track record of working collaboratively with clients on a range of digital initiatives, using an approach it calls co-creation. The follow-through, from starting at a business requirements level to tangible solutions delivery, is therefore very natural.

In conclusion

From consumer marketing and search automation to smart cars, smart factories and even smart cities, Deep Learning is coming of age.

From consumer marketing and search automation to smart cars, smart factories and even smart cities, deep learning is coming of age. When it comes to putting a DL system into production, the best practices for building or buying a platform are fundamentally the same as for HPC and Big Data: it needs to be reliable, scalable, agile, secure, stable, and easy to manage.

And because effective training requires large datasets and powerful computing capabilities, DL systems need to be able to handle these both effectively and relatively inexpensively. It can therefore draw upon the development and operations expertise amassed in both HPC and Big Data for its technology platform.

We hope our discussion in this paper will be useful to you as you continue to evolve your own AI agenda, and Deep Learning initiatives within that.

References and further reading

The following reports are available free of charge from www.freeformdynamics.com:

1. The Impact of Automation on IT Operations
   Are you ready for the software-defined datacentre?
2. The Enterprise Cloud Imperative
   Time to shake things up a bit?
3. Application Platforms Matter
   But how do you take the pain out of designing and building optimised systems?
4. Managing Cloud Complexity
   The emerging role of converged services

About Freeform Dynamics

Freeform Dynamics is an IT industry analyst firm. Through our research and insights, we aim to help busy IT and business professionals get up to speed on the latest technology developments and make better-informed investment decisions.

For more information, and access to our library of free research, please visit www.freeformdynamics.com or follow us on Twitter @FreeformCentral.

About Fujitsu

Fujitsu is the leading Japanese information and communication technology (ICT) company offering a full range of technology products, solutions and services. Approximately 155,000 Fujitsu people support customers in more than 100 countries. We use our experience and the power of ICT to shape the future of society with our customers.

For more information, please visit www.fujitsu.com.

Terms of Use