An InsightaaS/DC Foresight Community Report: June 2019

Industrialization of the Data Center: the Compute Factory

A DC FORESIGHT BEST PRACTICE REPORT

LEAD ANALYST: MARY ALLEN, INSIGHTAAS

PEER LEAD: FRANCOIS STERIN, CHIEF INDUSTRIAL OFFICER, OVH

CONTRIBUTING COMMUNITY MEMBERS: Kirby Peters, BMO Group, Mike Brown, TeraGo Networks, Chris Loken, Compute Ontario, Souvik Pal, Compute Infrastructure Research Centre (McMaster University), Ahsan Khan, ThinkOn Inc., Michael O'Neil, InsightaaS, Peter Near, VMware.

June 2019

OVH
Innovation for Freedom

DC Foresight

Contents

Industrialization of the Data Centre: The Compute Factory

Foreword

Research conducted by the DC Foresight community is designed to support development of best practices that accelerate the adoption and use of advanced technologies in Canada, and the operational benefits that are derived from these technologies. Each report is created with input from expert stakeholders, functioning as a working group, who bring a variety of perspectives to a multifaceted discussion of a core topic: here, the industrialization of compute service delivery.

Each DC Foresight report is unique, but all are based on a common framework which begins with a definition of the technology challenge, considers the business objectives (why would we invest in a solution?) that are driving interest in the topic, discusses best practices that address key aspects of the central issue, and identifies metrics and milestones that can be used by the technology user to calibrate deployment progress.

The *Industrialization of the Data Centre: the Compute Factory* initiative benefits from the insights shared by a working group comprised of IT leaders from a variety of service provider organizations (OVH, TeraGo Networks, ThinkOn); a large Canadian financial institution (the Bank of Montreal); and research organizations (Compute Infrastructure Research Centre at McMaster University, and Compute Ontario). The document owes a special debt of thanks to OVH Chief Industrial Officer Francois Sterin who acted as Peer Lead for this report.

Defining the issue: what is the compute factory?

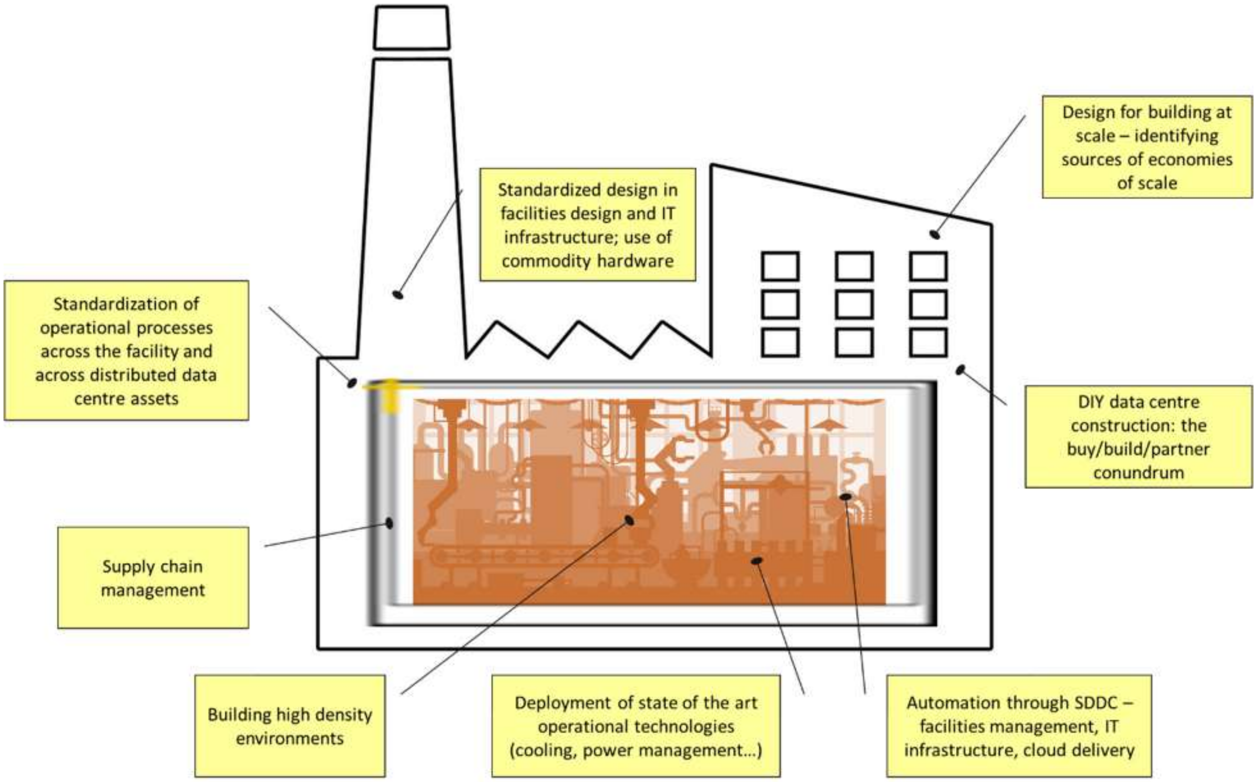
As it applied to the massive demographic and manufacturing transformation that marked modernization of economies the 19th century, the term ‘industrialization’ referred to new ways of organizing production. Fuelled by steam power and then electricity, technical innovations such as the cotton gin and mechanized looms combined with the centralization of labour in large facilities where operators applied scientific principles to the minimization of input and maximization of output. Over time, the assembly line, automation, and factory scale techniques evolved to help industrialists achieve economies of scale, a concept that outlined the cost advantage enterprises would obtain through expanding their operations and separation of tasks, first developed by Scottish economist and philosopher Adam Smith at the end of the 18th century. When applied to the world of IT, ‘industrialization’ describes similar trends and objectives: today, continuous tech innovation has combined with the centralization of compute resources, process automation, and the rapid scale-up of both compute output and productive capacity, which is controlled via sophisticated, software-based monitoring systems to minimize (power) waste and maximize efficiency.

The benefits of this hyperscale approach are broadly acknowledged. A large, highly automated, process-driven compute production facility that boasts strong performance metrics – the compute factory – is the aspirational goal of most data centre owners. What is less well-defined are its attributes: insight into what kinds of facilities are able to participate in this most recent industrial revolution.

While most IT (and business) professionals have an intuitive understanding of what industrialization in a data centre context means, specifics of the approaches taken to the design and execution of compute production taken by the hyperscale providers and others require further investigation. Understanding of the scale technologies and deployment practices that apply may support broader application of techniques aimed at improving data centre operational efficiencies, while helping consumers of compute resources better assess providers' service delivery capabilities. In this investigation, members of DC Foresight *Industrialization of the Data Centre* working group have considered the mechanics and the impact of the following attributes to the advance of industrialization in the data centre.

- Design for building at scale and identifying sources of economies of scale
- Standardized design in facilities systems and IT infrastructure - use of commodity hardware
- Standardization of operational processes across the facility, and across distributed data centre assets
- Supply chain management
- Building high density environments
- Deploying state of the art operational technologies (cooling, power management, etc.)
- Automation through software defined data centre capabilities – automation in facilities management, IT infrastructure and cloud service delivery
  DIY data centre construction – the buy, build, partner conundrum

Figure 1. The industrialized data centre: key considerations


Standardized design in facilities design and IT infrastructure; use of commodity hardware
Design for building at scale – identifying sources of economies of scale
Standardization of operational processes across the facility and across distributed data centre assets
DIY data centre construction: the buy/build/partner conundrum
Supply chain management
Building high density environments
Deployment of state of the art operational technologies (cooling, power management…)
Automation through SDDC – facilities management, IT infrastructure, cloud delivery

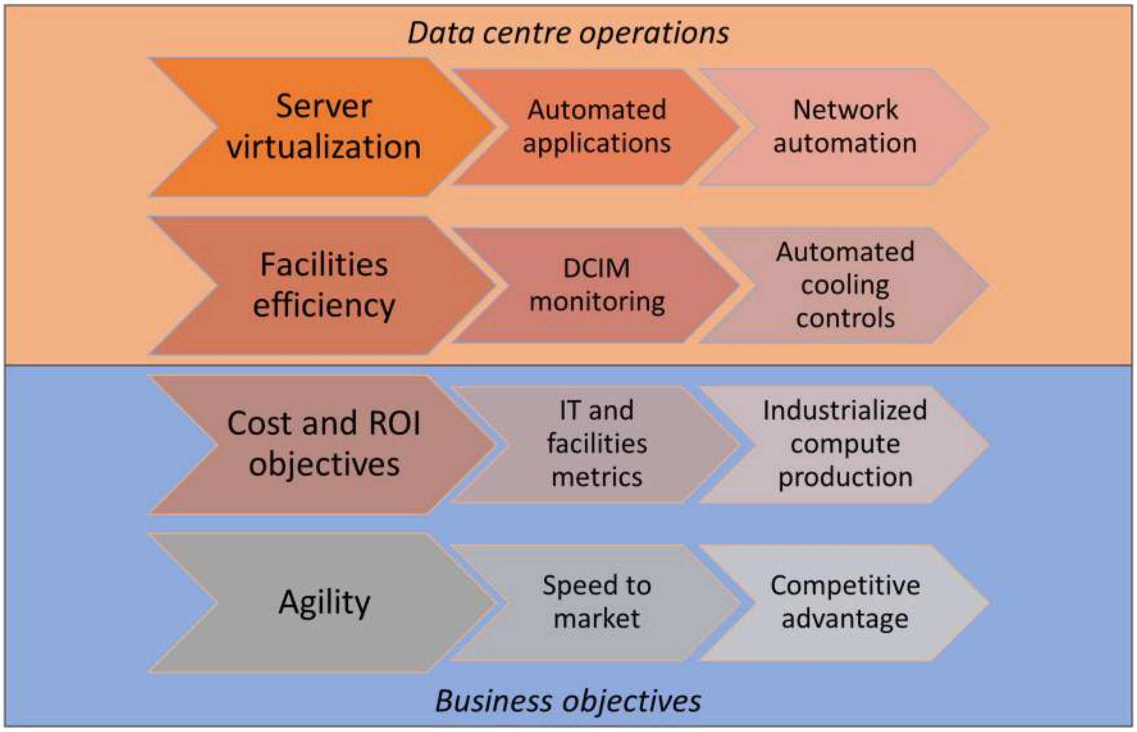
Source: DC Foresight/InsightaaS, 2019

Throughout history, industrialization has been closely associated with automation. The replacement of manual with automated processes in car manufacturing is an important example drawn from the North American experience. Henry Ford's first experiment with the assembly line – in production of the Model T's magneto ignition energy component – reduced time required for assembly of one unit by one worker from 15 minutes to approximately 7 per unit, through separation of tasks and delivery of work to workers by a moving line, a productivity improvement achieved in an experiment involving 29 workers. The benefits of Industrialization are not limited to large scale environments; automation can be applied to smaller environments and smaller processes, as in the Ford experiment, and so in IT. The key to achieving success lies in alignment of different practices with specific environments. In this report, members of the working group have outlined the benefits of industrialization of compute production, but also the application of discrete approaches to different kinds of compute facilities, including large hyperscale operations, enterprise data centre, research HPC environments, colocation facilities and the operations of smaller wholesale service providers.

Business objectives: customer, cost and connecting with the market

In data centre strategy aimed at the industrialization of compute resources, the automation of various processes can become a driver in itself. An organization may begin this journey by implementing server virtualization, which leads to an interest in automating applications, which in turn depends on automation of the network. The same progression may occur on the facilities side, where DCIM monitoring, for example, leads to implementation of automated control for cooling systems. This gradual evolution of automation works to build IT-as-an assembly line. But behind this process are business drivers that resonate with different levels of the user organization. At the C suite level, the goal of automation is agility, speed to market so that business groups can do things faster than competitors. In this scenario, compute just has to work without interruption and work quickly, a model that has been difficult to sustain as the increasingly complexity and volume of compute demand has put pressure on traditional data centre operations. At operational levels within the organization, cost and ROI become important input to efforts to industrialize compute production – the IT and facilities managers, for example, may be incented to demonstrate reduced cost of operation. But different business drivers are also a function of the type of organization that is looking to IT-as-an assembly line.

Figure 2. Parallel paths to industrialized operations

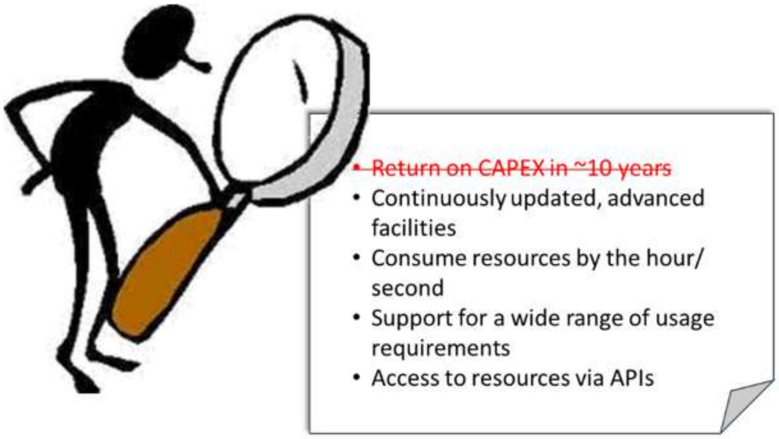


Source: DC Foresight/InsightaaS, 2019

Customer-driven innovation

Henry Ford is famously (if erroneously) credited with the observation that, if he asked people what they wanted, "they would have said faster horses," not cars. This perception of misinformed customer expectation rooted in historical models does not translate well to the contemporary consumer of data centre services. While a decade back, owners expected that a traditional data centre CAPEX build project would produce returns within 10 years, today, data centre customers want to consume resources by the hour, or even by the second, in advanced cloud facilities. To support this kind of user, data centres must build out platforms that are very reactive to different, and increasingly demanding customer usage requirements. This shift in the way consumers consume extends to information sharing: according to the working group, if consumers in the past wanted to understand provider capabilities and IP, buyers of compute resources now simply wish to access resources through an API, based on consume-as-you go business models.

Figure 3. Customer expectations ('not faster horses')



- ~~Return on CAPEX in ~10 years~~
- Continuously updated, advanced facilities
- Consume resources by the hour/ second
- Support for a wide range of usage requirements
- Access to resources via APIs

Source: DC Foresight/InsightaaS, 2019

For the provider of compute resources, these evolved expectations introduce new challenges to demand planning. IT and digital cycles are shorter, and customers' own need to move more quickly in competitive environments is placing new strains on delivery capability. According to the working group, without an agile model that can respond to varying demands, where capacity can be quickly reused, infrastructure quickly becomes inefficient. In this scenario, industrial innovation that delivers agility within IT and facilities infrastructure is now a requirement for addressing the needs of the new consumer.

Cost reduction

Another key driver of industrialization in the data centre is cost benefit, defined in terms of cost-effective design and planning, not low cost. While responsiveness – or the ability to react quickly to service demand noted above – is a critical characteristic of industrial IT platforms, the application of lean principles, optimization of all systems and sound business assessment of various deployment options can improve the cost outlook for the service provider, and ultimately, the end user. For example, in research environments that build HPC to support compute intensive workloads on defined operational budgets, the impact of economies of scale can be significant. To illustrate, the working group pointed to the importance of streamlining across systems: once bigger challenges, such as efficient whitespace management, have been addressed, economic benefit can be derived though optimization of systems and components. These improvements may have minor effects individually, but large aggregate impact. Smaller effects add up, the working group noted, when they are realized across millions of servers. To drive all value available in operational budgets, operators of research facilities have moved down the stack to consider opportunities to optimize at individual rack and server level with advanced control power, heating, and cooling systems.

An enterprise case illustrates the benefits of another industrialization tactic – separating tasks. The quick deployment of a separate, offsite powerhouse to support energy requirements proved to be a cost-effective alternative to building out power infrastructure within the data centre. Built in a controlled interior environment, rather than open site construction site, where exposure to Canadian winter climate conditions could affect quality, this approach enabled the enterprise to optimize use of space. While a stick build within the facility could involve the construction of empty space, since fees for the separate powerhouse were based on a per square foot cost, design was maximized to take advantage of every square inch. Building only what was needed to house three 2.7 MW generators, day tanks, and distribution gear produced cost savings over time, and the power solution enabled better scheduling of power supply, leading to better cost management, as well as better quality power, which delivers its own productivity benefits.

> In retrofit – or brownfield – environments, a modular approach to adding power capacity may be the only alternative – and in live environments, it is a preferred strategy as existing operations can remain intact: with proper planning, cutover involves no downtime

In retrofit – or brownfield – environments, this modular approach to adding power capacity may in fact be the only alternative. And in live environments, it is a preferred strategy as existing operations can remain intact: with proper planning, cut over involves no downtime, a more acceptable outcome in risk averse industries where service delivery interruption entails significant financial penalties.

Connecting with the market

In some instances, cost acts as a constraint on the ability to build according to certain principles of industrialization. For example, limited access to financial resources may preclude an organization's ability to construct a national footprint of large facilities, or to locate greenfield builds at sites with access to cheap power or better opportunities for free air cooling. In existing facilities, such as a multi-tenant colocation site, where operators have no control over what customers deploy in individual cabinets, creating cost efficiencies becomes especially problematic. But if participating in more obvious industrialization strategies is difficult for an existing, downtown colo, it is possible to develop techniques to optimize facilities systems on a daily basis, relying on the ability to react very quickly to customer requests. While cost benefit may be achieved through better power management, for example, ongoing opportunity may be realized by connecting with and responding to customers' changing need to monitor and reduce power consumption. Similarly, in wholesale environments, the ability to implement more quickly – execution in lockstep with demand – and a services orientation based on better management of IT and facilities can offer competitive advantage.

Figure 4. Industrialization barriers and enablers: colo facilities

Industrialization *barriers:*

- Access to financial resources
- Brownfield facilities
- Lack of control (colos)

Industrialization *enablers:*

- Providing better power monitoring and management
- Faster implementation (execution in lockstep with demand)
- Services orientation based on better IT and facility management

Source: DC Foresight/InsightaaS, 2019

Best Practices

The holy grail in data center operation is service reliability, a delivery state that the working group advises is achieved through advance planning that encompasses budget, customer expectations, design and technology options. This state is most easily reached in greenfield builds than in retrofit or upgrade of existing facilities that must drive out residual value in legacy systems. But what principles and practices inform planning for the compute factory? In its discussion of best practices, the working group has identified six key inputs that contribute to success in building scale environments.

Standardization – the cookie cutter to data centre design

In more purpose-built, traditional data centre models, the ability to tailor or customize service has been a delivery asset. In more modern approaches designed for rapid scale, standardization is a critical requirement. In mega-scale operations, ranging from large data centres that consume tens of megawatts of power to 150 megawatt facilities in China, standardization allows operators to maximize cooling efficiency, better distribute power, ensure power availability (through location close to grid power sources, and through generator capacity that delivers redundancy), and balance workloads across servers. Standardization, which enables replication that can drive down unit costs while supporting rapid replacement/repair in maintenance operations, applies to both facilities equipment and the IT equipment level. In many operations, the benefits of standardization are enhanced through the deployment of commodity hardware that is software defined and based on open standards systems. Today, there are several organizations working to drive standardization by easing deployment. The Open19 Foundation, for example, is looking to define a cross-industry common server form factor to support server 'plug and play' in data centres of all sizes, including edge deployments.

Standardization may also apply to operating procedures – regular, proactive maintenance routines or automated alert systems, for example, help operators avoid failure, and to more quickly mitigate outages that may occur. Standardization in maintenance and update processes is a best practice that is available to operations of all sizes.

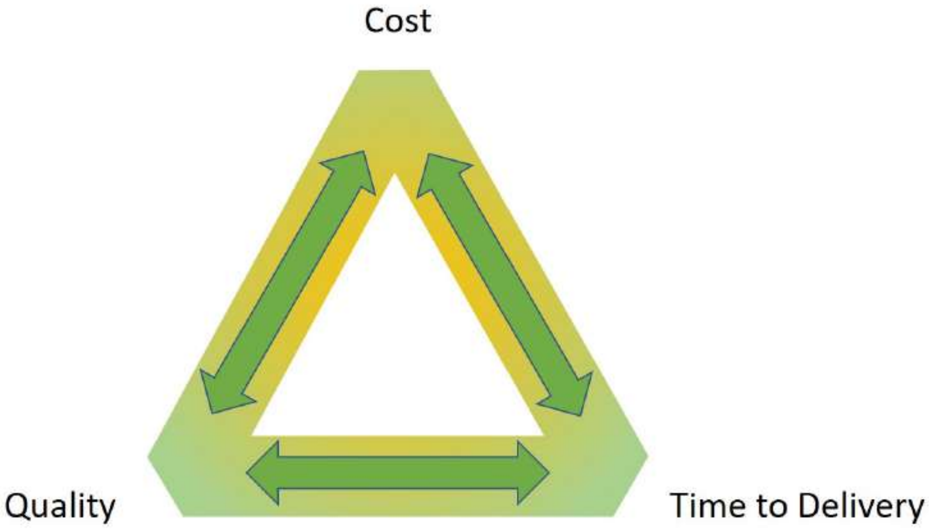
Supply chain management

As it simplifies equipment purchases, standardization also supports better supply chain management for systems optimization. In-house server assembly by the hyperscalers, for example, built from third party, standard processor components, can provide cost savings, and also address design requirements. In scale environments, tight integration of systems to optimize output is most readily achieved through control of the supply chain. In advanced operations, the goal is to plan for tight vertical integration throughout operations – from processor, to server, to rack, row and room. Mastery of the full supply chain also enables rapid provisioning – the operator knows what components are coming from the factory, and when they are available for connection into the data center – as well as greater agility in the introduction of new products. Cost modeling for a new CPU, GPU or server manufacture, for example, can help the data centre designer better assess the potential benefit in integrating specific new technologies.

In management of the supply chain, industrial sales and operations planning (S&OP) is a key enabler: to improve outcomes in S&OP, the industrialized data centre relies on sophisticated monitoring of both the facility and IT equipment to support Big Data analytics that can improve operational efficiencies, and also requirements forecasting for the server fleet. In today's on-demand infrastructure world, these kinds of tools and methodologies are increasingly important. While provisioning in a traditional facility may be viewed as 'IT project' where operators assess need on an ad hoc basis, the infrastructure supply chain enables a more agile 'Dev/Ops' approach that works to ensure continuous response to on-demand consumers of compute resources.

The ultimate goal in data-driven supply chain management is to balance needs in the 'cost-quality-time to deliver' triangle, where the impact of independent elements as well as the interdependencies between component cost, quality management, and time-to-deliver are taken into account to optimize operational output and address customer satisfaction.

Figure 5. Balancing key requirements in the standardized supply chain



Source: DC Foresight/InsightaaS, 2019

Building data centre density

The need to balance increasing server densities with cooling and space requirements is a familiar issue that is managed in hyperscale environments though the implementation of advanced technologies, such as precision cooling. In other kinds of environments, this balance has proven more elusive, and continues to dog efforts to industrialize data centre operations. In research computing, for example, where processing is concentrated, workloads cannot be stretched across the football fields of raised floor operated by the large service providers due to issues with communications speeds between racks and the cost for copper and fibre between them. Large bandwidth, software defined networking solutions may help to address this issue, as utilize virtualized connections to address physical communications challenges. However, ever higher density requirements for compute-intensive workloads have further stymied efforts to scale out. As the working group explained, while advances in the development of exascale systems have reduced calculations of power requirements from hundreds of megawatts a decade ago to 30 megawatts for a single system in the US, development of concentrated systems introduces significant new operational challenges from space and density perspectives. New exascale systems will likely occupy two or so racks, with cabinets that would be 300 kilowatts apiece. The need to deploy cooling at highly concentrated levels remains a concern in the HPC world.

Cooling models filter across data centre categories

According to the working group, in facilities with power densities of 20 or 30 kilowatts per rack, simple principles of physics preclude the use of air cooling techniques. As a result, HPC facilities, and many of the facilities used by hyperscale providers. have transitioned to the direct application of liquids to cool equipment. In other kinds of data centres – enterprise and colo, for example – many operators have inherited or continue to adopt architectures based on air cooling. As computing becomes more dense in these environments as well – an inevitable eventuality given new processor (especially GPU) capabilities – a manageable transition to liquid cooling architectures will become increasingly important. While the application of simple liquid/water cooling or even immersion cooling may be slightly different in traditional enterprise and colo facilities, heat exchange doors attached to individual racks makes these effectively autonomous – a configuration that might be applied to smaller facilities. Today, there are many industry groups working on the standardization of cooling distribution: the Open Compute Project, for example, is working on standards for water cooling. As these standards emerge, it is likely that concepts developed in larger environments will drive proposed solutions that are ultimately accepted across the rest of the market. However, density requirements are providing opportunity for new players, creators of off-the-rack liquid cooling solutions for GPU-intensive configurations and big memory CPUs, for example, to manage the inevitable increases in heat generation from new server classes.

Automation across IT and facilities fuels industrialization

In the delivery of compute services, automation is an increasingly critical input. For the large infrastructure as a service (IaaS) provider, automation ensures that systems and components which are introduced to the data centre immediately begin to generate revenue and services for the customer. For the smaller provider, automation supports a service delivery model in which it’s not necessary to be a giant facility in order to be cost-effective. Automation can drive efficiencies in small pilot systems that tap into emerging market opportunities, which are then scaled through the use of standardized hardware and other infrastructure.

An early use case is automated monitoring of different environmentals (heat, humidity) at the server or even workload level that has been used to track system health and to adjust heat and cooling in the facility, or to track PUE, a measure of energy efficiency in the data centre. But as the operator moves towards an industrialized model, automation extends beyond facilities equipment, to logical systems and the IT infrastructure. When applied to the compute side in a customer environment, automation operates at different levels: it works to manage existing compute infrastructure, and to manage access to additional resources, in cloud bursting models, for example, where it determines who is enabled, and how they get access to more machines. On the provider side, automation improves provisioning – the ‘rack and stack’ of the hardware to ensure compute is available – as well as on-demand delivery, which is enabled through the use of software thresholds that alert operators to the need for additional servers from the fleet to expand the resource pool, or to remove resources when a workload is scaled down or

> Service reliability, on which competitive positioning depends, hinges on the need to eliminate manual operation of the physical infrastructure, in favour of advanced monitoring and automation

stopped. In sophisticated provisioning systems, customers make requests via open APIs, and automated delivery of a compute rack generates the entire software stack that needs to be deployed on that rack or server.

In new design research, automation is also being applied to day-to-day operational tasks and processes to eliminate many of the human elements in infrastructure management. The goal is to develop routines based on computational modules that can track behaviors and detect what's normal and what's not normal. From a data centre management perspective, an abnormality can build, resulting in problems down the line, or subtle signatures can be detected, and issues addressed proactively. According to the working group, this kind of automated intelligence and insight into operating conditions will be key to the successful operation of increasingly massive data centre facilities. Service reliability, on which competitive positioning depends, hinges on the need to eliminate manual operation of the physical infrastructure, in favour of advanced monitoring and automation.

Other value-added data centre services are also enabled through automation. Off-hour support is one example; another is the automation of network connectivity. The latter may refer to automated provisioning of internal data centre networking, but more importantly, to the instantiation of access networks that include multiple providers. While manual connections into the data centre delivered by the network provider limit customers to a defined network, automated interconnection to multiple providers enables greater flexibility in a value-added, network-agnostic delivery approach.

The buy, build, partner conundrum – size matters

When set against the benefits of the industrialized model, the decision to build, buy or partner depends to a great extent on size of the operation. A data centre that has achieved volume levels (equipment, component input and delivery output) that enable economies of scale will be better positioned than a smaller operation to continue to build out its capacity. For this category of data centre, the instincts towards DIY which enable better control of vertical integration and cost savings may wane when a partner with specialist knowledge and its own scale operations can be engaged. For example, a third-party provider that can build racks more cheaply and delivery them more quickly at the same price points, while assuming some of the business risk in delivering this component, would be a good partner for many operators.

A data centre that has achieved volume levels (equipment, component input and delivery output) that enable economies of scale will be better positioned than a smaller operation to continue to build out its capacity

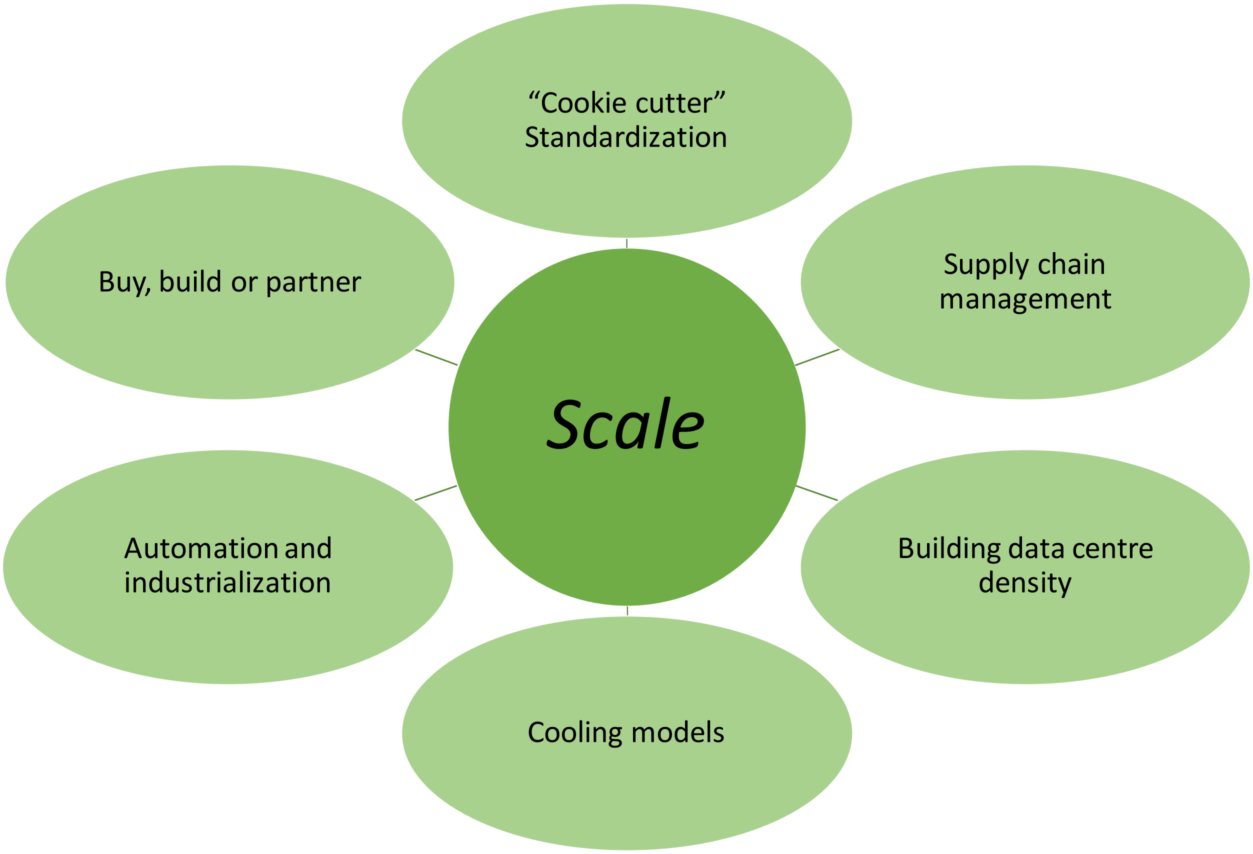
It is important to note that a desirable partner is not necessarily the lowest-cost commodity provider. Good partnerships are rooted in the ability to share development responsibilities in a long-term relationship, with both parties willing to co-innovate solutions that can be readily introduced to the data centre to support ongoing scale and innovation.

For the smaller cloud provider, partnership may prove to be a good option, as the capital outlay for building infrastructure can be significant and financial backing for a business that seeks to compete directly with the hyperscalers more scarce. A relationship with a partner who provides basic hardware/software on an as-needed basis in a profit-sharing arrangement may be a viable economic model. In a medium to large size enterprise, the buy/build discussion would depend on available capital, prioritization of OPEX to finance purchase of public cloud resources, and the availability and cost of specialized expertise to support advanced requirements such as security or AI provisioning in addition to automation of IT service delivery.

Key attributes of the compute factory – scale design, standardization, supply chain management with use of commodity hardware, advanced power and cooling to manage density requirements, and advanced automation – are in place to a greater or lesser degree in many service provider, and other large enterprise, operations. Other organizations that may not have the resources, the specialist expertise or the focus to implement these techniques, may take advantage of trends towards the industrialization of computing resources in a more indirect way. For example, the enterprise data centre may elect to shift certain workloads that are creating capacity challenges to the cloud; the efficiency gains wrested through deployment in highly industrialized environments should translate to cost advantages for the consumer of cloud resources.

In the HPC world, capability – defined as having thousands of cores and high-performance interconnects – was originally perceived as unavailable in the public cloud. In high utilization environments, such as the campus research compute system that is used 24/7 by researchers across the country, financial analysis shows that on-premise computing may be the most cost-effective approach. However, with improvements to cloud provider offerings – ex. 10 to 40G networking speeds – and more ready availability of components, such as the Google TPU, FPGAs, or neuromorphic chips that are used by very few individuals but critical to compute-intensive environments, the on-prem vs. cloud equations is beginning to shift. Today, many cloud providers offer GPU 'as a service' – a value proposition that is not replicable in low usage HPC or enterprise environments. The prospect of devoting more focus to value-adds that better align with specific research needs, such as developing efficiencies in research methods and coding and providing more extensive user support, is also driving more research computing into the commercial world. While there will always be a place in research computing for the creation of truly cutting edge advanced systems, the working group believes that a major share of compute loads could move to cloud today and that users would be happy for it – assuming they are able to maintain software, manage VMs and workloads, and work effectively within cloud provider business relationships.

Figure 6. Six keys to building scale


"Cookie cutter" Standardization
Buy, build or partner
Supply chain management
Scale
Automation and industrialization
Building data centre density
Cooling models

Source: DC Foresight/InsightaaS, 2019

Options for the middle road – dropping in modular capacity and the 3R advantage

For many organizations, data governance or other requirements rule out the shift to a complete outsourced model for the delivery of optimized cloud resources. But with access to advanced automation capabilities delivered in specialized products, enterprise, colo, or research facilities can provide competitive services in specific areas, or even approximate the efficiencies created by the large hyperscale operations. For example, by automating data centre management processes for small, remote facilities – potentially, consisting only of a couple of racks – where personnel are not available to perform maintenance functions needed to keep the operation running, the organization can deliver remote computing on-premise, rather than default to public cloud infrastructure.

'Plug and play' of different modules is another approach that can be used to access advanced industrial capabilities. In the enterprise powerhouse example noted above, an organization was able to quickly deploy a separate facility to supply energy to the main data centre that delivered cost efficiencies and also quality improvements. The use of shipping containers that house compute, storage, networking in converged infrastructure models, and pre-package all power and cooling requirements, is another way that the enterprise can quickly attain new levels of efficiency and scale. As the working group notes, a single shipping container could house thousands of CPUs; it would arrive on site pre-configured, pre-set up, ready to drop into virtually any location inside or outside the data centre to add capacity where and when it is needed. Connected to existing chilled water and power, the container could be up and running extremely quickly. For smaller data centres, this modular approach also provides cost advantages; the ability to start with a relatively small initial investment and build through incremental capacity as demand and revenue increase, enables many kinds of operations to participate in the industrialization of compute.

Through reduce, reuse and recycle, it is also possible to generate bottom line results for operators, customers, and the environment. Better life cycle management of IT assets, for example, allows data center operators to reduce the number of servers that enter landfill, to refurbish some for reuse – second and third life for servers can strengthen business models – as does disciplined recycling which delivers cost back from jobbers, while ensuring responsible end of life disposal for assets. As it reduces environmental impact, this triple R approach improves cost outlook for providers and for their customers – those, for example, with less critical workloads who are willing to deploy on refurbished machines. Better lifecycle management can also be extended to facilities – in data centre upgrades, for example, new types of racks can be used, and older ones retrofitted to address the need for improved power density. When data centre refresh is not restricted by legacy infrastructure, the operation has made good progress in the industrialization of infrastructure. In larger facilities, the cost/benefit equation for this approach may be more compelling; however, lifecycle management may be deployed effectively in all kinds of facilities for cost, customer and environmental benefit.

Metrics – key indicators of the progress of industrialization

Provider focus on utilization and customer satisfaction

In the data centre industry, several sets of metrics have evolved to allow operators, and consumers of data centre services, to judge the performance of facilities assets and the delivery of IT services. The goal of measurement is to set benchmarks, compare current with past state and with performance indicators for competitive operations, and to respond to these results with techniques and/or technologies designed to improve operational efficiencies, energy consumption/carbon emissions, and service performance. These metrics may also act as a proxy indicator of progress towards industrialization, which ultimately works towards the same outcomes.

> According to the working group, the ‘80/20 rule’ applies to service delivery: 80% of demand will be standard work that can be delivered right away, and the more industrialized an operation is, the more quickly a standard product can be delivered to the customer

PUE, or power utilization effectiveness, is the most commonly used metric. A ratio that describes the total amount of energy used by the data centre facility divided by the energy used by the IT equipment, PUE was developed by The Green Grid consortium, and accepted as an ISO/IEC standard in 2016. For many years, it has served as a benchmark for data centre operators to assess the energy efficiency of their facilities, and with it, other measures such as carbon impact. While an imperfect measure, the introduction of PUE has served to galvanize power saving activity across the data centre industry. PUE is now used in market outreach as a positioning statement on an operators' relative efficiency, innovation, and environmental responsibility. As a response to the increasing carbon footprint of data centres, and associated energy costs, this activity is having an impact. A recent Lawrence Berkeley Lab study has shown that since 2016, overall energy consumption in the US industry has levelled – an indication of improved energy performance since the number and size of large facilities have grown to service increasing demand. Report authors attribute improvements to better cooling techniques, power scaling (scale back of energy consumption for idle servers), and consolidation, including the use of cloud technologies and a shift from distributed to hyperscale computing.

Interestingly, the last two energy saving techniques identified by the LBL are associated with management of IT infrastructure, as opposed to facilities equipment. A key criticism of PUE as an efficiency measure has been neglect of potential opportunities for improved efficiencies on the IT side of the house. Other evaluations have focused on the lack of standardization in applying PUE: it can be measured in multiple ways, at different times of the year (temperature can have an impact); factors such as oversubscription rates are not included, and there is no consistency in the way PUE numbers are communicated. This critique of PUE has led to the creations of other measures of data centre consumption, including CUE (which measures carbon emissions), WUE (water utilization), the ITEE (IT equipment efficiency) or ITUE (which measures utilization efficiency per IT load). Another metric that is gaining traction is the BEF, or base energy factor, which addresses difficulties in comparing 'like to like' in PUE by establishing a baseline representing the optimal utilization for best energy performance that a facility can achieve. This baseline becomes the basis for all performance assessments and comparisons.

From an industrialization perspective, higher resource utilization, achieved through better capacity planning is an important success metric. In some cases, this is achieved via consolidation, which is in turn driven by deployment of higher-capacity (and often, higher density) IT systems: 10x reductions in racks needed to support a particular workload, and corresponding savings in space and power, are not unheard of in industrialized environments. Data centre operators will also quantify industrialization progress through additional measures, such as economies of scale which dictate that unit cost per component decreases as volume grows – and size of the operation increases. A final category of metric seeks to improve customer satisfaction through reduced lead time for delivery to customers, including specific requests. According to the working group, the 80/20 rule applies to service delivery. Though 20% of requirement will involve specific needs that are more complex to address, 80% of demand will be standard work that can be delivered right away: the more industrialized an operation, the more quickly a standard product can be delivered to the customer.

As an important input to industrialization, automation carries its own set of metrics that speak to the efficiency of operations, service delivery performance and business outcomes for the internal/external customer. At the software level, the number of manual tasks that have been automated in areas that people don't want to work in, or in which human intervention is inefficient – the movement of workloads, provisioning of workloads, changing firewall rules, or the movement of applications within the environment, for example – can produce ROI associated with headcount reduction, as well as productivity benefits through reassignment of staff to higher value activities. And while cloud automation benefits can be measured in the number of firmware updates that are avoided (and less time in diagnostics when something goes wrong), automated ROI tools can show the return on better capacity management and from hardware and power optimization.

Building satisfaction – the customer view

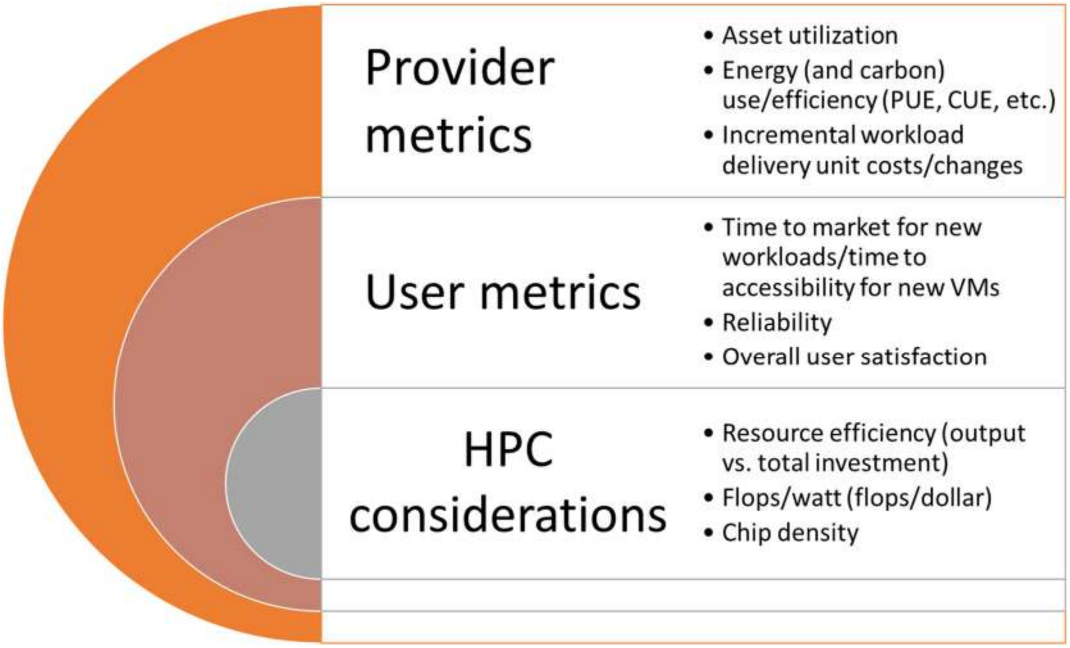
At the client level, the impact of industrial automation can be evaluated by time-to-market (or even in time to VMs), which is the time required for a project to deliver business benefit. A measure of speed in service demand fulfillment, this metric translates to how quickly data centre operators can ready infrastructure for business users or the development team, allowing consumers of data centre services to evaluate a provider's ability to support their goals. Speed is not the only criteria used by consumers to assess service performance though. Cost per machine is another indicator of service value, which may be used to compare efficiencies achieved in cloud vs. highly automated on-premise environments, or the relative efficiencies of different cloud offerings.

Reliability is another important priority for consumers that is measured through a variety of metrics. In messaging to potential customers, operational specs typically include built in redundancy (factor of N+1, etc.) as an indication of service reliability. For customers, a key evaluation metric is the rate of abnormal incidences that have occurred in cloud and colocation facilities, and how these are tied to outages. Another issue for buyers is the number of security breaches that have occurred, which may impact operations or put customer data at risk. To address this issue – as well as the compliance audit for certain workloads that a hoster may have to undergo – many providers have now implemented automated monitoring systems for visibility into security attacks.

In some industries – particularly in HPC environments common to the research world (and increasingly, used to support advanced workloads like AI as well), performance metrics are translated into measures of productivity. For example, in many research environments, utilization levels are a critical measure of success; however, utilization translates not necessarily to the share of time a server is used, but rather to the efficiency with which users take advantage of compute resources. The goal is to maximize scientific productivity per dollar in an environment that has finite resources, to ensure that a defined number of CPUs are used in as productive manner as is possible to maximize scientific output. To illustrate, the working group pointed to inefficient code that might consume a huge amount of RAM and only one core – blocking resources that might be used by another researcher. It's possible to monitor this kind of activity by placing performance counters on the nodes that can feed back measures of the number of flops that are being used, memory bandwidth utilization or disk utilization, which are then correlated with jobs or users to identify potential productivity gaps or inefficient users of the compute resources.

Other examples from the HPC community include flops/watt (which eventually translate to flops/dollar); organizations pursuing this metric favour extensive use of GPUs. Management in these environments is also likely to view watts/chip as a meaningful metric: many facilities were built when 85 watts per chip was standard, but today's high performance systems regularly weigh in at 150 watts per chip, and specialized chips (like GPUs) currently run as high as 300 watts per chip, which has clear implications for power distribution and cooling strategies. Per-chip metrics are becoming important in commercial environments as well, with workloads like AI able to use – and benefitting from – 200-300 watt chips. Conventional wisdom dictates that at these densities, water cooling becomes an important element of the industrialized data centre approach.

Figure 7. Key metrics used to gauge industrialization and facility performance



Provider metrics

- Asset utilization
- Energy (and carbon) use/efficiency (PUE, CUE, etc.)
- Incremental workload delivery unit costs/changes

User metrics

- Time to market for new workloads/time to accessibility for new VMs
- Reliability
- Overall user satisfaction

HPC considerations

- Resource efficiency (output vs. total investment)
- Flops/watt (flops/dollar)
- Chip density

Source: DC Foresight/InsightaaS, 2019

This focus on productivity is an approach that users across industries can take to assess value obtained through data centre services. By applying a scientific lens that will instrument efficiencies, and the cost and business value achieved through use of cloud, on-premise or hybrid resources, users replicate the approach taken by advanced, industrialized providers of services who design and model to maximize efficiencies, hearkening back to the scientific principles that early economists argued should ultimately govern the input/output equation.

Sponsoring members and contributors

Industrialization of the Data Centre – the Compute Factory was sponsored by the OVH Group, and is the work product of the DC Foresight best practice community. In addition to OVH, key sponsoring members of the DC Foresight community include Vertiv, Lenovo, the Technology Integration Group, ThinkOn, Cologix, Cisco, Intel, Belden and AMD. The DC Foresight gratefully acknowledges the support of these corporate members:

Contributors to this document

We would like to acknowledge the following contributing DC Foresight community members, whose input shaped this document.

Kirby Peters

Director, Critical Facilities, BMO Financial Group

A mechanical engineer with broad experience in vendor and enterprise communities, Kirby now delivers global planning/operational oversight for BMO critical facilities, working with bank business leaders to understand needs and mitigate risk.

Mike C. Brown

Manager, Eastern Engineering Team, TeraGo Networks

With responsibility for strategy and technical operations, Mike manages TeraGo's data centres and fibre optics networks. Before TeraGo, he spent over a decade in management roles in engineering and operations for multiple for tier-1 companies.

Chris Loken

Chief Technology Officer, Compute Ontario

A CTO with deep experience in procuring, supporting and supporting HPC environments in the research sector, Chris now works with Compute Ontario to ensure researchers have the data centre resources needed to further their research

Souvik Pal

Principal Research Engineer, CIRC at McMaster University

A postdoctoral fellow in McMaster University's Department of Mechanical Engineering, Souvik specializes in thermal management. He also serves as principal research engineer at McMaster's Computing Infrastructure Research Centre.

Ahsan Khan

Chief Technology Officer, ThinkOn Inc.

A solution architect with strength in storage and virtualization technologies, Ahsan now acts as CTO at Canada's only IaaS wholesaler, leading development and implementation of the company's technology strategy.

Michael O'Neil

Principal Analyst, InsightaaS

Canada's leading IT industry analyst, Michael has helped executives at leading organizations capitalize on new technologies and business opportunities. He has authored hundreds of reports, and three acclaimed books on cloud and analytics.

Peter Near

National Director, Solution Engineering, VMware Canada

Peter has finetuned solution engineering consulting with key Ontario software firms. He currently leads a pre-sales solution team at VMware focused on helping clients deploy advanced technology to ignite digital transformation.

Peer Lead: Francois Sterin, Chief Industrial Officer, OVH Group

Francois has brought over a decade's worth of experience as head of the data centre and energy portfolio for an iconic US hyperscaler to the OVH Group, the leading European provider of hosting and data centre services. Technically a unicorn, OVH has experienced rapid growth, based on global expansion of its data centre footprint and development of cloud services. Francois has been a key contributor to the OVH growth strategy, leading development of the company's technology strategy and overseeing engineering activities to ensure reliability and scalability of the company's service delivery infrastructure.

Lead Analyst: Mary Allen, Chief Content Officer, InsightaaS

Co-founder of InsightaaS and the Toronto Cloud Business Coalition, Mary has spent two decades understanding and communicating key trends that shape Canadian and global IT markets. As journalist and analyst, she has authored hundreds of articles, reports and analyses on advanced technology deployment, including (with InsightaaS partner Michael O'Neil) the acclaimed management book, *Building Cloud Value: A Guide to Best Practice, 2016*.

Mary still likes Green IT.

About the OVH Group

Headquartered in Roubaix, France, OVH is a privately held French cloud computing company that offers VPS, dedicated servers, cloud and other web services. OVH was founded 20 years ago by current chairman Octave Klaba, as a web hosting firm. Since 1999, the company has experienced rapid growth in terms of product offerings and data centre inventory: OVH now boasts 27 data centres, which host 300,000 servers, in 19 countries, including Canada. The OVH data centre in Beauharnois outside Montreal is a model of industrial construction and operation that exhibits many of the scale attributes discussed in this report.

OVH positions as the "alternative cloud', and is looking to compete with the large US hyperscale service providers. To support this ambitious goal, the company bought VMware's former vCloud Air assets in 2017 as an entre to American markets, and continues to innovate not only in the build of infrastructure assets but in an expanding portfolio of open source product offerings.

About InsightaaS

Dedicated to exploring "the 'why' in enterprise technology," InsightaaS was founded by Mary Allen and Michael O'Neil in 2013. The company operates Canada's deepest IT content website and provides strategic consulting and channel management guidance to leading firms in Canada, the US and abroad.

In 2015, InsightaaS launched the Toronto Cloud Business Coalition, a community dedicated to the co-creation of Best Practice guidance designed to accelerate adoption and use of cloud in Canada. The tremendous success of the group has spawned four additional communities – IoT Coalition Canada, Canadian Analytics Business Community, V2V: The Economics of Data, and DC Foresight – and several meetup groups, notably CIA-Plus. These groups continue to help Canadian businesses to capture value from advanced technology.