Bloor InDetail

InDetail Paper by Bloor
Authors **Philip Howard**
Publish date **May 2018**

IBM Cloud Private for Data

> We would especially like to emphasise that one of the difficulties we have seen with companies trying to implement AI is a disconnect between data scientists who develop the relevant models, and those responsible for deploying those models in production.

Executive summary

Artificial Intelligence (AI) and machine learning (a subset of AI) are the topics du jour within not just the IT and analytics communities but for business in general. They are constantly being referenced across the media and within boardrooms. There is general agreement that there is huge potential value in implementing the techniques and technologies associated with AI and machine learning, both directly in terms of increased efficiency and as a competitive differentiator. Not to mention the ability to introduce new business models and services. Nevertheless, while it may be easy to see the potential benefits, implementing them is not so simple. Forecasts suggest – see ***Figure 1*** – that very large numbers of companies will be investing in AI over the years to come but, today, only a relative handful have been able to do so. There are many reasons for this: cultural change issues, security concerns, talent acquisition and others, which we are not in a position to discuss. However, there is one issue – limited or no technological capability with respect to data and analytics – that we will consider in this paper and, specifically, we will discuss IBM Cloud Private for Data.

IBM Cloud Private for Data is an integrated data science, data engineering and app building platform built on top of IBM Cloud Private (ICP). The latter is intended to **a)** provide all the benefits of cloud computing but inside your firewall and **b)** provide a stepping-stone, should you want one, to broader (public) cloud deployments. Further, ICP has a micro-services architecture, which has additional benefits, which we will discuss. Going beyond this, ICP for Data itself is intended to provide an environment that will make it easier to implement data-driven processes and operations and, more particularly, to support both the development of AI and machine learning capabilities, and their deployment. This last point is important because there can easily be a disconnect between data scientists (who often work for business departments) and the people (usually IT) who need to operationalise the work of those data scientists.

During the course of this paper we will first discuss ICP and then ICP for Data. However, it is worth discussing, briefly, some general reasons why you might want to deploy either of these. In the case of ICP there are a couple of reasons that go beyond the flexibility and scale provided by cloud computing in general. The first is the ability to modernise enterprise applications by refactoring these, using the microservices offered by ICP. And, secondly, there is the ability to build cloud-native applications which may either be new or which leverage existing applications and data, and perhaps including public cloud services, while keeping your data securely behind your firewall. Many of the use cases for ICP for Data are really an extension of these capabilities, allowing operational applications to have decisioning or machine learning built into them: trained within ICP for Data but deployed either directly or via a cloud-native application, as you prefer. And there other use cases: for example, you could easily deploy a data lake (or lakes) using ICP for Data with a data catalogue and data federation (see below) spanning all of these, so that they do not become siloed.

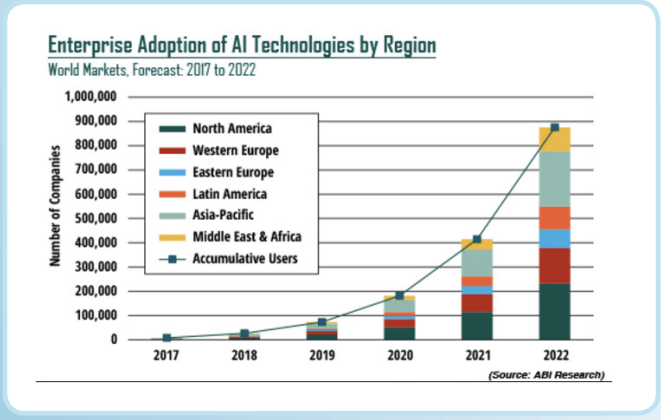

Figure 1: Forecast adoption of AI

However, it is not our intention to write reams about the features of these products nor to go into details about use case. Rather, we intend to reference the business benefits these offerings are intended to address, as well as to discuss the issues that can arise within this environment. We conclude with a section on IBM's short-term roadmap for new releases of ICP and ICP for Data.

> **"There is huge potential value in implementing the techniques and technologies associated with AI and machine learning, both directly in terms of increased efficiency and as a competitive differentiator."**

IBM Cloud Private

> **This significantly speeds up release cycles (continuous delivery) and enables incremental improvements to applications.**

ICP is a private cloud. Note that this is distinct from a virtual private cloud where a public cloud vendor acts as the service provider, with ICP either your IT department acts as the service provider or you can get a third-party to manage it for you, but in either case business units act as tenants. The basic principle is that you can get the benefits of a cloud deployment but with your data remaining protected behind your firewall, though it can be hosted externally, if you prefer. ICP interoperates with traditional on-premises deployments as well as both public and virtual public clouds, so it supports hybrid environments and is seen by IBM as a potential stepping stone to public clouds in the future.

The virtues of cloud-based computing, particularly the advantages to be derived from elastic scaling, rapid deployment, scalability and so forth, are well-known and we do not intend to rehearse those issues here. However, ICP (and ICP for Data) has a microservices architecture, as illustrated in ***Figure 2***, and it is worth discussing the benefits that this brings, as the use of microservices, based on Docker and Kubernetes, is a relatively new concept.

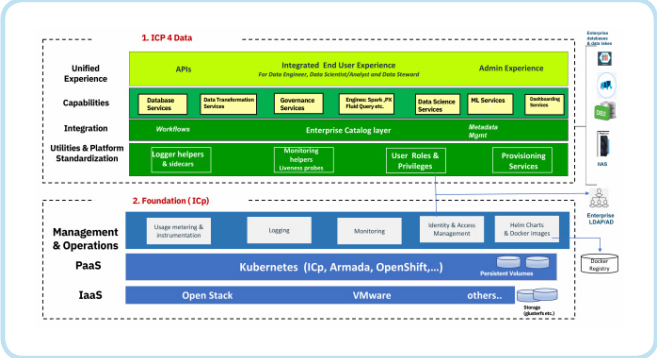

Figure 2: ICP for Data architecture

According to ***www.microservices.io*** *"microservices – also known as the microservice architecture – is an architectural style that structures an application as a collection of loosely coupled services, which implement business capabilities. The microservice architecture enables the continuous delivery/deployment of large, complex applications."* Microsoft goes further when it says (amongst other things) that *"in some ways, microservices are the natural evolution of service-oriented architectures (SOA) but there are differences. Some of the defining characteristics of a microservice include the fact that services are small, independent and loosely coupled; that each service is a separate codebase, which can be managed by a small development team; that services can be deployed independently and that an existing service can be updated without rebuilding and redeploying the entire application; and that services communicate with each other using well-defined APIs with implementation details hidden from other services."*

What IBM has done with ICP and ICP for Data is to take its existing capabilities and recreate them – where that makes sense – as services, as shown in ***Figure 2***. Some of the advantages of this approach are described in the quotes above, but there are others. For example, we should note that this encourages interoperability. It makes it much easier to take a database service and combine it with a machine learning service, for example, because you can pick and choose what services you want to use. The other point that we should emphasise, is that this significantly speeds up release cycles (continuous delivery) and enables incremental improvements to applications. This is important, not just because you get new features faster, but also because traditional release cycles of a year or eighteen months, are usually disruptive and result in delays to new releases being implemented. Moreover, bearing in mind that this is a cloud-based environment, this means that IBM can also incrementally introduce new capabilities in the same fashion.

IBM Cloud Private for Data

Supporting AI initiatives requires that you obtain all the data you need, govern it to ensure it is trustworthy, analyse and build the machine learning and other algorithms necessary for the project in hand and, finally, to be able to put the results of this exercise into production. This is not a trivial task. The individuals and groups – shown in ***Figure 3***, which illustrates the components of ICP for Data – responsible for these activities are often disparate and disconnected and it requires a collaborative approach to make this work efficiently. Moreover, it needs a set of capabilities that are beyond most software providers and, to enable the sort of collaboration required, you would really prefer a consistent user interface across all the underlying software capabilities that are necessary. While we will discuss ICP for Data in more detail shortly, one of the most notable features of this new product introduction is precisely that it provides a common user experience across its software stack. And it is worth commenting that it is the micro services architecture that underlies ICP for Data that has enabled this, by decoupling the user interface from the individual software components, and then creating a new consistent interface. For reference, ***Figure 4*** provides a couple of screenshots that illustrate this interface.

Referring back to ***Figure 3*** there are a number of elements of the ICP for Data architecture that require discussion. Fundamental is the Enterprise Data Catalog. Essentially, this is like a library catalogue in that it holds details of all the data assets that are available to the organisation. However, unlike a library which simply gives you a reference for each named book, the data catalogue allows you to search by category so that you can, for example, find all assets relating to sales, customers or products. In other words, it allows you to find assets of value to your role that you might not otherwise be aware of.

Above the Enterprise Data Catalog in ***Figure 3***, are three boxes. IBM refers to these as *"collect, organize and analyze"*. If it were down to us we would probably put data integration into the *"collect"* box and rename *"organize"* as *"governance"* but that is merely a semantic quibble.

The *"collect"* box in principle supports the ability to leverage all sorts of data sources, which is obviously necessary: if you don't have the data you can't analyse it. However, this element of ICP for Data is relatively limited – compared to *"organize"* and *"analyze"* – in this release (it will be significantly extended in the next version: see the roadmap section below) and only supports the various elements of the Db2 family (Db2 itself, Db2 Warehouse, Db2 Event Store and so on) as well as Hadoop (Hortonworks, with which IBM has a partnership) and Big SQL (IBM's SQL on Hadoop engine). Attention should be paid to the fact that Db2 Event Store is supported, because this will be especially important in Internet of Things (IoT) based environments where you need to store data originating in sensors, actuators and other edge devices. Also provided within the *"collect"* box is data federation, which allows distributed queries across third party databases such as Oracle. However, for the moment this is data federation across relational databases not data virtualisation across both structured and unstructured data sources (see roadmap).

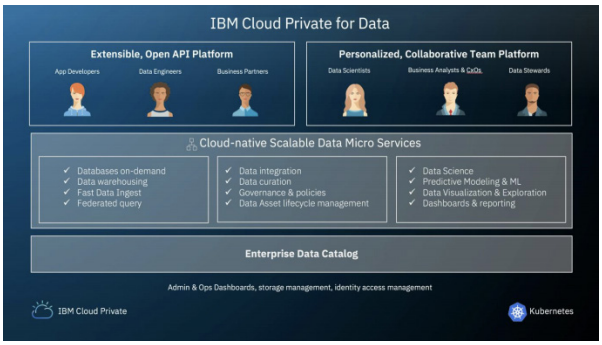

Figure 3:
ICP for Data components

Both the *"organize"* and *"analyze"* capabilities are more complete in this release than the *"collect"* services. As far as *"organize"* is concerned, ICP for Data leverages familiar technologies from the InfoSphere brand, with support for data cleansing, data masking, governance and so forth. In this context, it is worth pointing out that there are implications with respect to AI and machine learning that may not be immediately obvious. For example, consider automated decisioning. That is, where some sort of computerised algorithm

> **"There are implications with respect to AI and machine learning that may not be immediately obvious."**

makes decisions automatically, as a part of a relevant process, based on information received. That algorithm has to have absolutely trustworthy and unbiased data to work with, otherwise its decisions may be flawed, which could have very adverse consequences for your business. In other words, data quality – often thought of as just about names and address deduplication and cleansing – is fundamentally important within this context, and applies not just to structured data but also to semi-structured and unstructured data such as sensor readings and text. Note too, that in IoT environments there can be issues such as out of sequence, missing and duplicated readings, each of which must be catered for. In addition, you can get situations such as *"sensor drift"* where maximum and minimum sensor readings gradually increase or decrease (typically, because of extreme environmental conditions) and anomalous spikes (was this an event of interest or a loose connection?).

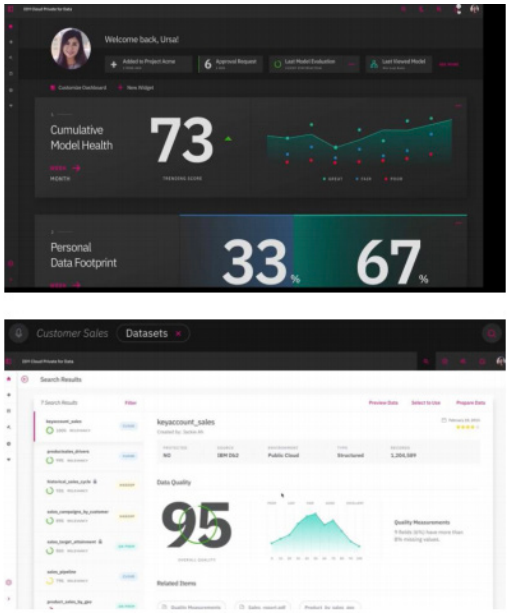

Figure 5: The coomon user experience of ICP for Data

The *"analyze"* box similarly leverages existing IBM technologies, including SPSS, Cognos and Watson Studio, which includes IBM Data Science Experience, which has specific capabilities for training, persisting and scoring machine learning models. However, there is a gap between trustworthy data and machine learning models. Data from different sources needs to be put into a consistent format, may need to be transformed (many algorithms work on value ranges between zero and one, which also explains why null values are an issue) and will need to be joined into a single dataset before you can start to run analytics against that data, or create machine learning algorithms.

Where we do take issue with IBM on ***Figure 3*** is the division between the top two boxes. We understand that this is a marketecture diagram, but this does not accurately reflect IBM's intentions or capabilities. It suggests, particularly, that App Developers and Data Engineers do not share the *"personalised, collaborative team platform"* – which includes crowd sourced recommendations as well as more prosaic capabilities such as workflow – available to other users. Our understanding is that this is inaccurate. As, indeed, it should be. We would especially like to emphasise that one of the difficulties we have seen with companies trying to implement AI is a disconnect between data scientists who develop the relevant models, and those responsible for deploying those models in production. And, if data engineers are defined as those people who prepare data ready for analytics, while the data scientists actually do the science, then this applies to them too. So, what is required is *"AnalyticOps"* – analogous to DevOps – in the sense of bridging the gap between the analytics and their operational deployment, which is precisely what the collaborative environment in ICP for data is intended to provide. And, we should add, being able to provide collaboration across this entire spectrum of roles and capabilities is only possible because of the breadth of IBM's capabilities.

"What is required is *"AnalyticOps"* – analogous to DevOps – in the sense of bridging the gap between the analytics and their operational deployment, which is precisely what the collaborative environment in ICP for data is intended to provide."

In this context we should also add that model management, while not explicitly shown in ***Figure 3***, is supported by ICP for Data. This is important because what is the best model today may not be the best model next year. This is for two reasons: firstly, next year you will have more real data to work with and, secondly, because conditions and trends change over time. This means that you may need to replace one algorithm or model with another periodically. Ideally, this should be on a *"hot swap"* basis with no downtime. In any case, the performance of models needs to be monitored and changed when appropriate, hence the need for model management.

Finally, note that ICP for Data leverages all the underlying capabilities of ICP itself, for administration, security, logging, monitoring and so forth.

Roadmap

As we write ICP for Data Enterprise edition is currently available in beta and is slated to be released shortly (May 29th 2018). As can be seen in ***Figure 5***, the Community Edition and a Cloud Native Edition, are scheduled for release during the second half of 2018 (probably earlier rather than later). However, there are also substantial upgrades to the Enterprise Edition planned. To begin with, the first release runs on Linux on x86 platforms, but IBM intends to extend this to other platforms. Further, from a cloud perspective, ICP for Data runs on IBM SoftLayer only at present but, again, the company intends to make it available on other cloud platforms in the future. In this context it is worth noting that IBM and Red Hat have announced that ICP and ICP for Data will be available on the Red Hat OpenShift Container Platform as Red Hat Certified Containers. IBM has also announced that ICP and ICP for Data will be ported to the AWS and Azure cloud platforms. Finally, IBM also plans to make ICP for Data available as an appliance.

As far as individual features are concerned, the most notable will be the extension to the capabilities of the *"collect"* set of components. In particular, data federation will be replaced (or supplemented) by data virtualisation. And, in general, ICP for Data will support a much broader range of both structured and unstructured data sources.

It is worth once more commenting that the rapid introduction of the various roadmap features described, and others we have not mentioned, are enabled because of the microservices architecture. Without that you could be waiting until next year before getting this additional functionality.

IBM has also announced that ICP and ICP for Data will be ported to the AWS and Azure cloud platforms.

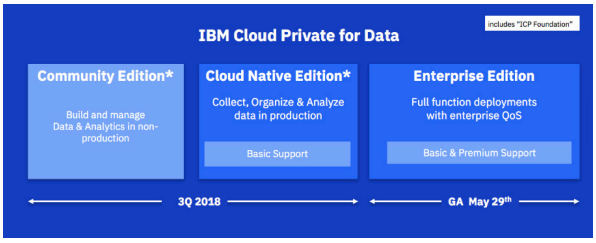

Figure 5: ICP for Data editions

Conclusion

At the Data Works summit in Berlin in April 2018 a straw poll was taken of the audience, asking how many of the attendees' companies planned to put data and analytics into the cloud. A, perhaps surprising, 34% of the 400+ people who voted, said their companies had no such plans. The truth is that no matter how impressive the hype is, there are many organisations that are reluctant to take that step, for a variety of reasons. That does not mean that they do not recognise the benefits of cloud-based computing but, currently, it is perceived to be a step too far. What ICP for Data offers is an in-between position: the benefits of cloud computing without the risk of moving data outside of your firewall.

However, this isn't all that ICP for Data offers: if you want to deploy machine learning – and almost everybody does – then you need an environment that facilitates that. IBM refers to this by saying that you can't have artificial intelligence without an information architecture (*"AI requires IA"*). And the problem with building an information architecture is that it involves many moving parts, many software requirements and many personas. To make this work requires that companies adopt AnalyticOps as a principle, and this requires not just a broad range of base functionality but collaborative support across all of the personas involved. Even though ICP for Data is in its first incarnation you can see that this is the direction in which the product is headed. It would be infinitely harder to achieve with a set of disparate products from multiple vendors.

Even though ICP for Data is in its first incarnation you can see that this [supporting AnalyticOps] is the direction in which the product is headed. It would be infinitely harder to achieve with a set of disparate products from multiple vendors.

FURTHER INFORMATION

Further information about this subject is available from
www.BloorResearch.com/update/xxxx

About the author

PHILIP HOWARD
Research Director / Information Management

Philip started in the computer industry way back in 1973 and has variously worked as a systems analyst, programmer and salesperson, as well as in marketing and product management, for a variety of companies including GEC Marconi, GPT, Philips Data Systems, Raytheon and NCR.

After a quarter of a century of not being his own boss Philip set up his own company in 1992 and his first client was Bloor Research (then ButlerBloor), with Philip working for the company as an associate analyst. His relationship with Bloor Research has continued since that time and he is now Research Director, focused on Information Management.

Information management includes anything that refers to the management, movement, governance and storage of data, as well as access to and analysis of that data. It involves diverse technologies that include (but are not limited to) databases and data warehousing, data integration, data quality, master data management, data governance, data migration, metadata management, and data preparation and analytics.

In addition to the numerous reports Philip has written on behalf of Bloor Research, Philip also contributes regularly to ***IT-Director.com*** and ***IT-Analysis.com*** and was previously editor of both ***Application Development News*** and ***Operating System News*** on behalf of Cambridge Market Intelligence (CMI). He has also contributed to various magazines and written a number of reports published by companies such as CMI and The Financial Times. Philip speaks regularly at conferences and other events throughout Europe and North America.

Away from work, Philip's primary leisure activities are canal boats, skiing, playing Bridge (at which he is a Life Master), and dining out.

Bloor overview

Technology is enabling rapid business evolution. The opportunities are immense but if you do not adapt then you will not survive. So in the age of Mutable business Evolution is Essential to your success.

We'll show you the future and help you deliver it.

Bloor brings fresh technological thinking to help you navigate complex business situations, converting challenges into new opportunities for real growth, profitability and impact.

We provide actionable strategic insight through our innovative independent technology research, advisory and consulting services. We assist companies throughout their transformation journeys to stay relevant, bringing fresh thinking to complex business situations and turning challenges into new opportunities for real growth and profitability.

For over 25 years, Bloor has assisted companies to intelligently evolve: by embracing technology to adjust their strategies and achieve the best possible outcomes. At Bloor, we will help you challenge assumptions to consistently improve and succeed.

Copyright and disclaimer

Bloor Research International Ltd
20–22 Wenlock Road
LONDON N1 7GU
United Kingdom

Tel: **+44 (0)20 7043 9750**
Web: **www.Bloorresearch.com**
email: **info@Bloor.eu**