InDetail Paper by Bloor
Authors Philip Howard
Publish date November 2019

IBM Cloud Pak for Data

Email
murat@accuras.com

Website
www.accuras.com

Phone
647-878-8463

Address
51 Gould Lane
TORONTO, ON L4J9B5 CANADA

> 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 Pak for Data. This was initially released in May 2018 (as IBM Cloud Private for Data), but this paper reflects the most recent release of the product (version 2.5) as of October 2019. The product is currently available in two versions: a Cloud Native Edition (supporting up to 64 virtual cores) and an Enterprise Edition (unlimited virtual cores). There is also IBM Cloud Pak Experiences, which allows you to test drive the environment without installation or cost, as well as a Quickstart for AWS public cloud customers, which provides a 90-day free trial of the software.

> 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 Pak for Data is an integrated data science, data engineering and app building platform built on top of Red Hat OpenShift Container Platform. The intention is 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, it has a micro- services architecture, which has additional benefits, which we will discuss. Going beyond this, IBM Cloud Pak for Data 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 Executive summary 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 consider the underlying platform and then its hyper-converged deployment option, IBM Cloud Pak for Data System, specifically. However, it is worth discussing, briefly, some general reasons why you might want to deploy either of these. In the case of Red Hat OpenShift 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. 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 IBM Cloud Pak for Data are really an extension of these capabilities, allowing operational applications to have decisioning or machine learning built into them: trained within IBM Cloud Pak for Data but deployed either directly or via a cloud-native application, as you prefer. And there are other use cases: for example, you could easily deploy a data lake (or lakes) using IBM Cloud Pak for Data System with a data catalogue and data virtualisation (see below) spanning all of these, so that they do not become siloed.

Enterprise Adoption of AI Technologies by Region / World Markets, Forecast: 2017 to 2020

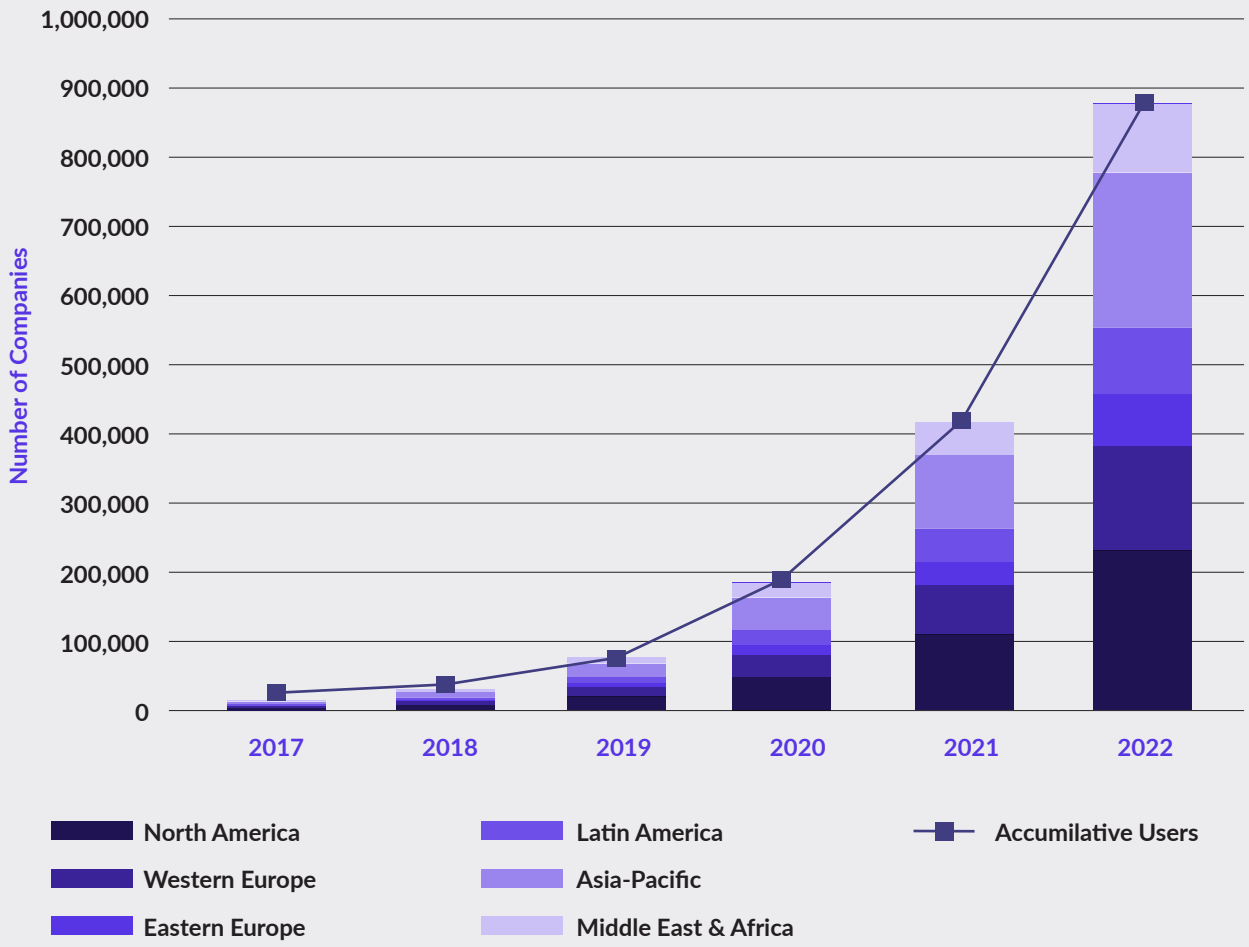

Figure 1: Forecast adoption of AI

However, it is not our intention to write reams about the features of these products nor to go into details about use cases. 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.

IBM Cloud Paks

T here are actually five IBM Cloud Paks at the time of writing: IBM Cloud Pak for Data, IBM Cloud Pak for Applications, IBM Cloud Pak for Automation, IBM Cloud Pak for Integration and IBM Cloud Pak for Multi-cloud Management. And there are two additional deployment options: IBM Cloud Pak for Data System and IBM Cloud Pak for Application Workloads. What all of these have in common is that they are based on modular container software that has been designed to support cloud-based environments, whether in building on, moving to, or managing that environment. All of these "paks" are portable and can run on-premises or in public clouds, they have been pre-integrated and are based on a combination of IBM and open source software, all of which is certified by IBM to provide full stack (including hardware) support.

It is not our intention to provide a detailed review of these various IBM Cloud Pak offerings, nor of the Red Hat OpenShift platform upon which they are based. However, a general discussion will be appropriate. In practice, IBM Cloud Pak offerings will often be deployed as a private cloud. Note that this is distinct from a virtual private cloud where a public cloud vendor acts as the service provider, here 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. The software 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. IBM Cloud Pak for Data System leverages all the underlying capabilities of IBM Cloud Pak for Data itself, for administration, security, logging, monitoring and so forth.

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

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, IBM Cloud Pak for Data (and IBM Cloud Pak for Data System) has a microservices architecture, 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.

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*

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.

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."

One further point to note is that IBM has announced the integration of Kubernetes with Cloud Foundry. While not yet available as a part of IBM Cloud Pak for Data this is significant because it will make Kubernetes deployment much simpler.

What IBM has done with IBM Cloud Pak for Data and IBM Cloud Pak for Data System is to take its existing capabilities and recreate them – where that makes sense – as services. Some of the advantages of this approach are described in the quotes above, but there are others. For example, we would 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.

Hyper-converged appliance

With IBM Cloud Pak for Data System, IBM delivers the core capabilities of the IBM Cloud Pak for Data platform on "hyper-converged" infrastructure. What IBM means by this is that all the necessary hardware, networking and software is pre-assembled for you with the company claiming a typical deployment time of just four hours. This is not technically an appliance because the hardware is installed within existing racks (assuming you have some free). Moreover, it is not a fixed appliance in the sense of limiting your use of other software, as it allows further capabilities to be implemented, as required, in a plug-and-play fashion. It is more of a pre-installed package of hardware and software rather than a traditional appliance. A starting configuration begins with 8 nodes, 128 cores and 1.5 Terabytes of storage. Compute and storage can be scaled independently. If you add nodes then this fact is automatically recognised by the management software. There is also an option to use FPGAs (field programmable gate arrays) for hardware acceleration. According to IBM these are four times faster than GPUs (graphical processing units), are half the price and use one eighth of the power.

The company also plans to make IBM Cloud Pak for Data available on the Power and Z mainframe platforms.

IBM Cloud Pak for Data System

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 2***, which illustrates the components of IBM Cloud Pak for Data – responsible for these activities are often disparate and disconnected and it requires a collaborative approach to make this work efficiently. Moreover, it requires 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 IBM Cloud Pak for Data in more detail shortly, one of its most notable features of this product 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 IBM Cloud Pak for Data that has enabled this, by decoupling the user interface from the individual software components, and then creating a new consistent interface, a number of examples of which are included in this paper.

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

There are a number of elements of the IBM Cloud Pak for Data architecture that require discussion. Fundamental is the Enterprise Data Catalog, which forms a part of the Data and Analytics Governance layer. 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. This is illustrated in ***Figure 3***.

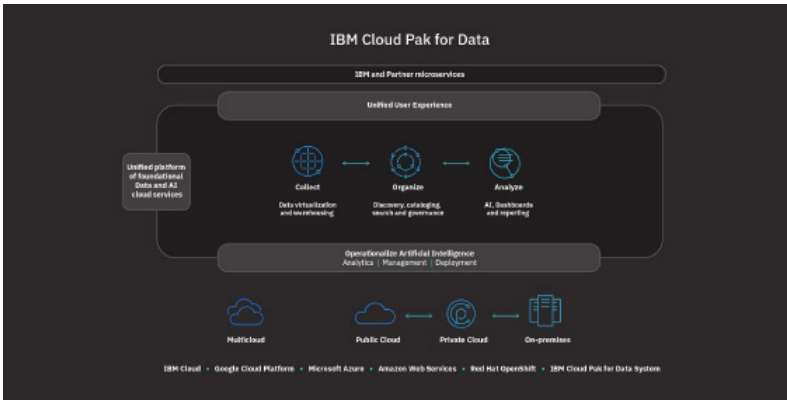

Figure 2: The components of IBM Cloud Pak for Data

Where we take issue with IBM on ***Figure 2*** 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 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 it should be.

Above the Data and Analytics Governance layer in ***Figure 2***, are four boxes. IBM refers to these as "collect, organize, analyze and infuse". 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. We will discuss each of these separately.

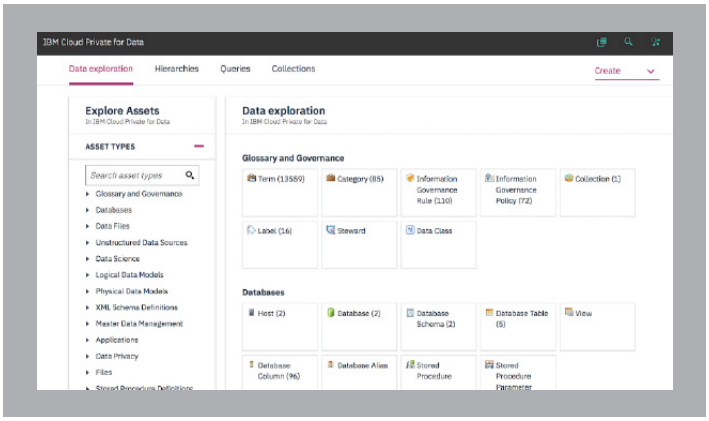

Figure 3: Exploring data assets via the Enterprise Data Catalog

Collecting data

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. There are two principal aspects to this: what is listed as "databases on demand" and "data virtualization", which need separate discussions. In the latest release the former covers the entirety of the Db2 family (Db2 itself, Db2 Warehouse, Db2 Event Store and so on) as well as Hadoop, Big SQL (IBM's SQL on Hadoop engine) and MongoDB (a new addition in this release). 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.

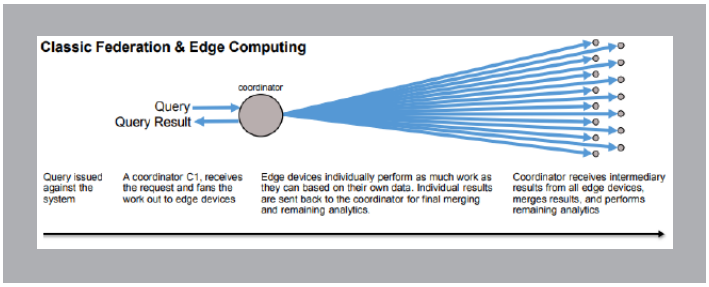

Figure 4: How traditional data federation works versus how IBM Data Virtualization works

When IBM Cloud Pak for Data was initially released the "collect" box included data federation but this has been replaced with data virtualization. This represents a significant enhancement to the product and it is worth describing how it works because IBM's capabilities in this area are both innovative and ahead of other offerings in the market. To understand how this technology (which was previously codenamed QueryPlex) works, it is worth comparing how it differs from conventional approaches to data federation, shown in ***Figure 4***, with IBM's use of a computational mesh, illustrated in ***Figure 5***.

As can be seen, the key difference is that the computational mesh not only performs analytics locally but also within a local constellation. Given that moving the data across the network is the biggest issue with traditional data federation techniques this should significantly improve performance. Moreover, this should have a significant impact

on costs because you need less infrastructure to get the same (or better) performance. Data sources supported by the computational mesh include the Db2 family (as above), Netezza, BigSQL, Informix, Derby, Oracle, SQL Server, MySQL, PostgreSQL, Hive and Impala. Further support is in the pipeline and is discussed in the Roadmap section below. Note that Data Virtualization integrates with the Enterprise Data Catalog.

While ***Figure 5*** gives a conceptual view of how IBM Data Virtualization works, what the user sees is illustrated in ***Figure 6***. In other words, as is described in ***Figure 5***, the edge devices organise themselves and you don't need to know how that is done. Note that although this refers to "edge devices", which would be entirely appropriate for IoT environments, it also applies to databases and other sources that need not necessarily be on the edge.

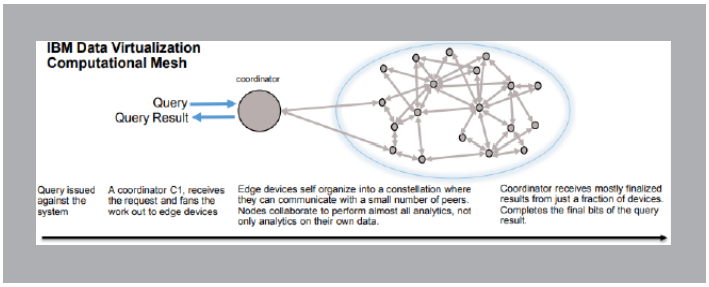

Figure 5: How traditional data federation works versus how IBM Data Virtualization works

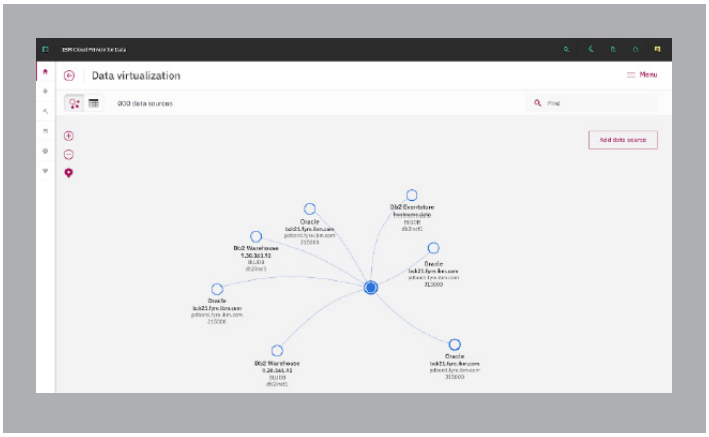

Figure 6: The user view of IBM Data Virtualization

Organising and analysing data

As far as "organise" is concerned, IBM Cloud Pak 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 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?).

With respect to analysis capabilities IBM Cloud Pak for Data leverages existing IBM technologies, including SPSS Modeler, Watson Explorer and Watson Studio, which has specific capabilities for training, persisting and scoring machine learning models. IBM Cloud Pak for Data also includes support for IBM ILOG CPLEX Optimization Studio.

This is a prescriptive analytics solution that enables the development and deployment of decision optimisation models using mathematical and constraint processing. Watson Assistant, which is used for building Chatbots, is also available within IBM Cloud Pak for Data as a premium add-on.

Finally, we should also add that model management is supported by IBM Cloud Pak for Data through SPSS. 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.

Infusing data

The concept behind "infuse" is that data and analytics are all very well but unless their use is infused throughout the organisation then you will not be able to maximise the benefits of that data and those analyses. This is especially true when it comes to AI. 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 IBM Cloud Pak for Data System 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. To further enable this, IBM Watson OpenScale has been built on top of IBM Cloud Pak for Data and is available as part of the base in the latest release.

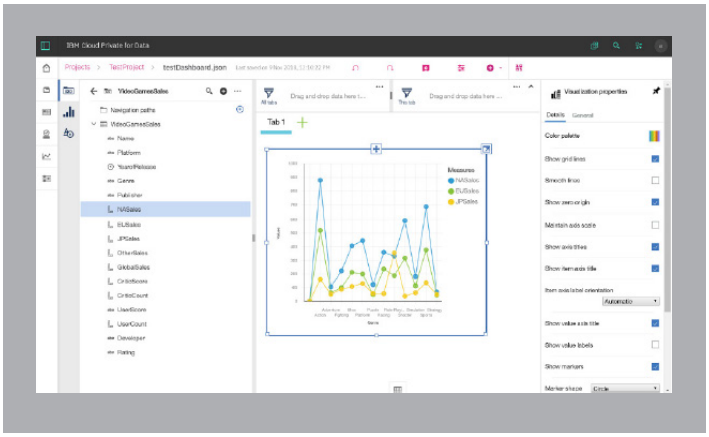

Figure 7: Drag and drop interface for data visualisation in ICP for Data System

Watson OpenScale has various capabilities that enable business to operate and automate AI across its lifecycle, with the aim of bringing enterprise-scale AJ more easily within reach. First, the product allows you to monitor how the deployment of AI is affecting business outcomes, whether positively or negatively, no matter where models were developed or deployed.

Figure 8 shows a dashboard that is used to automatically detect unfair or inaccurate results based on user-defined conditions. As can be seen in this example, "driver performance" has two issues: it is both inaccurate (that is, performing badly) and is producing unfair outcomes. In total there are three accuracy alerts and six bias alerts.

A second function of Watson OpenScale is to detect, and automatically mitigate, bias: see *Figure 9*. Various companies have announced such capabilities for training purposes but, as far as we know, Watson OpenScale is the only such product that is also available at run-time, when it compares the performance of the deployed models with recommended de-biased models. No retraining is required. However, it will provide a way for a user to download and inspect the de-biased model before deploying it into production.

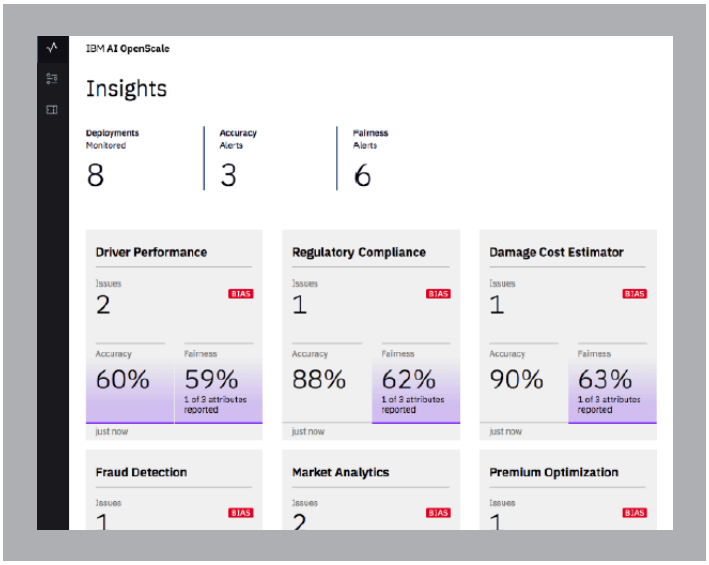

Figure 8: Acuracy and Fairness in Watson OpenScale

Thirdly, in addition to audit capabilities that allow you to trace the recommendation behind any decision, Watson OpenScale provides an explainability function that allows you to explore the factors involved in any decision, how much those factors contributed to the decision and how different factors might have changed the outcome. Confidence levels are provided for the transaction in question, and there are facilities to support the business in explaining a decision to a customer, regulatory body or other party.

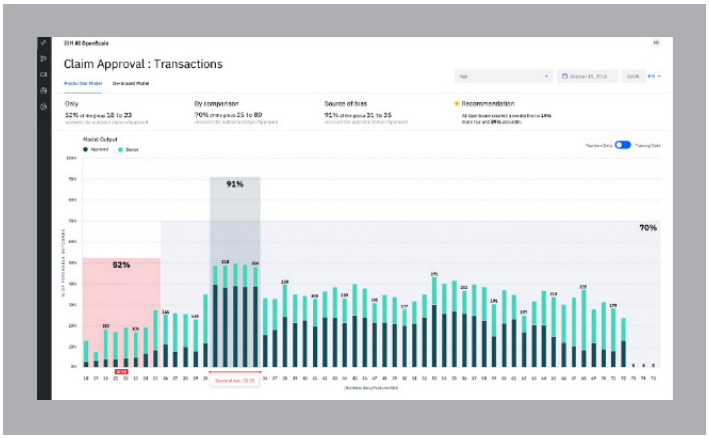

Figure 9: Bias detection in Watson OpenScale

Under the covers the product also provides what is known as payload logging. This means that all the data being scored is logged and persisted to a database and it is against this data that the various dashboards and other capabilities of Watson OpenScale are based. You can query the database using IBM Cognos or using third party tools, in order to get performance metrics. In addition, these metrics can also be exported and combined with application metrics to measure business outcomes.

Finally, Watson OpenScale automates other aspects of the AI lifecycle through NeuNets (Neural Network Synthesis). This is a function that will automatically create and recommend customised neural networks for AI deployments in any runtime environment, creating significant productivity gains for the data science teams tasked with building and retraining AI.

Recent updates

There are three notable updates that are worth mentioning.

The first is the Unified Console, which is used to manage the "Databases on Demand" feature of IBM Cloud Pak for Data. Significant enhancements to this are planned. Secondly, Cognos Analytics is now supported as part of the product's "analysing data" capabilities, as is Watson Discovery and, indeed, the entirety of the Watson portfolio as premium add-ons. And, thirdly, Data Virtualization includes support for Excel, CSV and text files, MongoDB, SAP HANA, SAS, MariaDB, CouchDB, Cloudant, various Amazon and Azure databases, sundry streaming products, multiple mainframe environments (IBM and others), generic JDBC access and several third-party data warehousing databases.

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 IBM Cloud Pak 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 IBM Cloud Pak 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 ("no AI without 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 IBM Cloud Pak for Data is still developing 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 IBM Cloud Pak for Data is still developing 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.

Next steps

For more information, visit IBM Cloud Pak for Data page.

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 system 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.