Latest weather forecast: more clouds on the horizon...

Cloud collaboration has allowed projects of previously unimaginable scale and scope to be constructed at an unmet price/performance ratio. Widespread access to Information and Communication Technologies (ICTs) and agile and adaptable systems of workflow now allow people across the globe to collaborate on virtual mega-engineering projects that are unprecedented in scale or scope[1].

Adopting new computer and infrastructure technologies in the drug discovery and development community only makes sense if it results in better and faster scientific research. IT buzz words are often overwhelming scientists without translating these great technology innovations into viable scientific solutions. Life Sciences research is becoming increasingly collaborative and complex, leveraging multiple technologies to get a systems level understanding of diseases and organisms. This exponential increase in the scale of data being generated combined with increased collaboration has resulted in a need to rethink how data is cost effectively stored, analysed and shared among teams. Cloud computing is a combination of technologies and service offerings and has the potential to increase the speed of basic research projects significantly. Gartner, a leading information technology research and advisory company, predicts that the worldwide market for cloud services will grow from $58.6 billion in 2009 to $148.8 billion by 2014.

What are the benefits, negatives and opportunities of the use of cloud computing? The majority of publications about cloud computing in pharmaceutical discovery are product or technology centric. Would it not be more appropriate to start with the end user in mind and look at the application from a user-centric perspective?

Democratising research

The cloud as an infrastructure gives researchers computational access on a subscription or pay-by-demand cost structure. Instant access to computational power, significantly lower administration costs, no capital spending headaches and no dependence of availability of IT resources. Software as a service (SaaS) is a software delivery model in which software and its associated data are hosted centrally in the cloud and are typically accessed by users using a thin client computer or tablet device, using a web browser over the Internet. Zero footprint applications imply that no software needs to be pre-installed on your client. This significantly helps to simplify installation procedures. A browser is all you need! To upload successfully large datasets, it is critical to have access to fast network infrastructures. Limited network bandwidth, especially in start-up phase, may result in frustration and should be avoided.

As a researcher you want to avoid thinking in computer terms. **Table 1** summarises the overall nomenclature and major acronyms for the most common cloud computing service models: SaaS, PaaS, and IaaS.

Basic research in most academic disciplines is undergoing a fundamental shift from the three traditional paradigms of theory, experiment and computation to a new fourth paradigm of data-driven
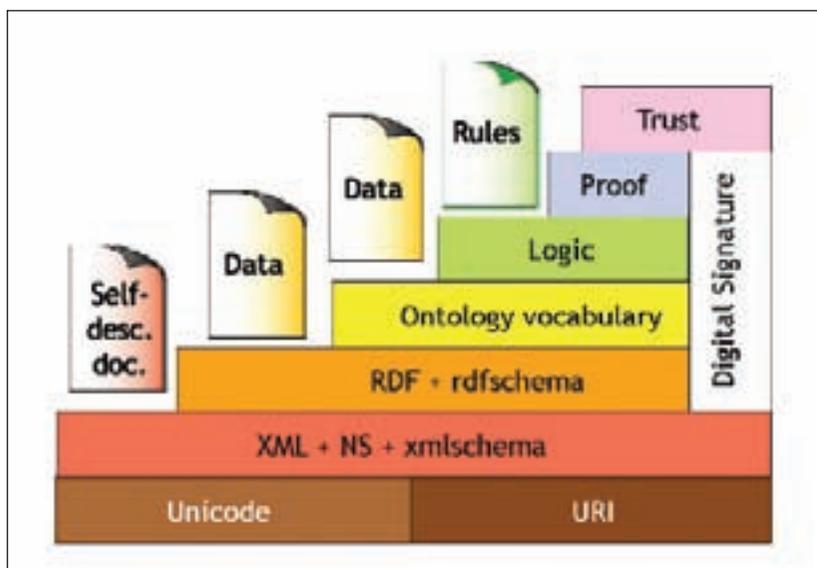
By Peter Boogaard

Informatics

Table 1: Cloud 101 abbreviation list

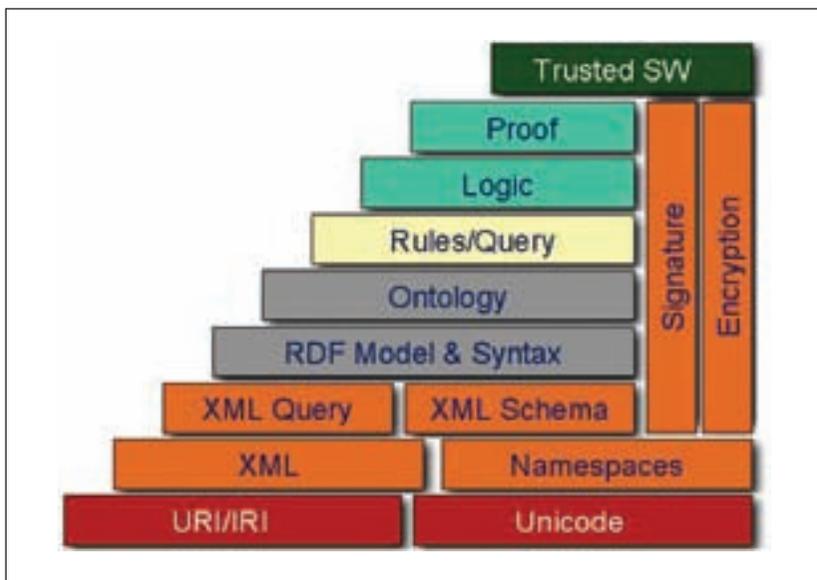|		
SaaS On-demand software	Software as a Service – Software distribution model in which discreet applications are hosted by a vendor or service provider and made available to customers over the Internet	• Easier administration • Automatic updates and patch management • Compatibility and easier collaboration: All users will have the same version of software • All operational costs. No capital investment. Pay as you go and subscription based
PaaS STaaS	Platform/STorage as a Service – External service which provides the hardware, operating system, software upgrades, security and everything else related to the day to day hosting of an (enterprise) application	• Unification of programming development efforts • Enabling geographically distributed development teams • Lower operational costs; cross-boundary, single vendor support • Enterprise data integration and systems consolidation
IaaS HaaS	Infrastructure as a Service – Provision model to outsource equipment used to support operations, including storage, hardware, servers and networking components	• Utility computing service and billing model • Automation of administrative tasks • Dynamic scaling • Desktop virtualisation • Policy-based services
DaaS Virtual desktop Hosted desktop	Desktop as a service – The outsourcing of a virtual desktop infrastructure to a third party service provider	Service provider manages the back-end responsibilities of data storage, backup, security and upgrades • All operational costs
XaaS Anything as a service Everything as a service	Refers to an increasing number of services that are delivered over the Internet rather than provided locally or on-site	XaaS is the essence of cloud computing
Public Cloud	Standard cloud computing model available to the general public over the Internet	• Inexpensive. Hardware, application costs are covered by the provider • Scalable to meet needs • No wasted resources; you pay for what you use • Examples include Gmail, Hotmail, Yahoo, SalesForce.Com
Private cloud Internal cloud Corporate cloud	Marketing term for a proprietary computing architecture that provides hosted services behind a firewall	• Full control over data • Internally hosted • Traditional approach
Hybrid cloud	A hybrid cloud is a composition of at least one private cloud and at least one public cloud	Allows a business to take advantage of the scalability and cost-effectiveness that a public cloud computing environment offers without exposing mission-critical applications and data to third-party vulnerabilities
Multi-tenancy architecture	An architecture in which a single instance of a software application serves multiple customers. Each customer is called a tenant. Tenants may be given the ability to customise some parts of the application, such as colour of the user interface (UI) or business rules, but they cannot customise the application's code	• Cost effective – provider only has to make updates once • Ability to take advantage of virtualisation and remote access • High security – each tenant's data is isolated and remains invisible to other tenants

discovery. "ICT democratises innovation and enables smaller companies, academia and students to get access to computing that were previously only accessible by large organisations," said Rüdiger Dorn, Director Cloud Computing, Public Sector EMEA of Microsoft Corporation. "Pay by

Informatics

Semantic technologies allow different conceptual domains to work together as a network

a High Performance Computing infrastructure has been significantly changed from weeks to minutes, and doesn't require IT staffing anymore, since it is all offered as an external service. From a financial perspective the model is attractive. Instead of spending six or more digits capital investment dollars after a lengthy budget and planning process, you now align your budget using operational costs with the pay-as-you-go pricing model and stop paying for equipment that sits idle between experiments. The cloud has the potential to be the best new development going mainstream for scientific researchers. Access to an almost unlimited amount of computer power to achieve scientific calculations and text searches, combined with an almost unlimited size of disk space for an affordable price. If you want to read more about new innovations in cloud computing, visit Amazon's cloud website at http://aws.amazon.com/what-is-aws/.

Licensing software in the cloud is different than traditional software licensing. A traditional software purchase involves, in many cases, a capital investment and comes with an annual maintenance and support plan. The SaaS model is solely based upon an operational cost model with monthly or annual subscription fees with all maintenance and support services included. No hidden costs. No surprises.

What's in it for me?
Now we can focus on how these new technology breakthroughs can help to increase discovery research, create more evidence on these new unlimited discoveries and creating new evidence for new inventions. Industrial Lab Automation observed the market and summarised the major trends.

The two major areas where cloud services are being deployed in discovery are computational power to execute computer intensive calculations and in electronic lab notebooks (ELN) to electronically capture scientific experiments and share knowledge across research teams.

John McCarthy, VP, Product Management of Accelrys, explains that its customers are successfully using Pipeline Pilot with the Next Generation Sequencing Collection with the Amazon EC2 web service for its next generation sequencing. "We can spin up any number of nodes in the cloud cluster and perform high intensity computational jobs and analyse terabytes of sequencing data to support our clients' work either internally or with their CROs in India and China."

Research is an extremely information-intensive endeavour. As a rule, experimental data are both dynamic and distributed. New data and properties

the cycle, without computational infrastructure, significantly accelerates new innovations." The Azure Research Engagement project aims to change the paradigm for scholarly and scientific research by extending the power of the computer into the cloud (http://research.microsoft.com/en-us/projects/azure/).

Amazon Elastic Compute Cloud (Amazon EC2) is a web service that provides resizable computing capacity in the cloud. It is designed to make web-scale computing easier for developers. It is a popular service to create scalable and highly available IT infrastructures which can store, compute and share multiple terabytes of data. The effort to build such

of research samples and models are continuously invented. It becomes very difficult to track the entire experiment from hypothesis and methods through to the raw data, to the processed and analysed data to the results to the conclusions to the reports and then to the management decision processes to find and use the information.

As data become distributed into cloud repositories it becomes increasingly important to use semantic methods to annotate and be able to relate these data to other scientific information. This is essential for many reasons. If scientists cannot easily find information to use or to share, there is little value to collecting it in the first place. Readily searchable repositories support collaboration and minimise unintentional duplication of effort. Links also allow researchers to communicate about data without having to send copies of it around by email – and multiple users can create annotations that point back to a common data object even if stored in a cloud repository. "The use of ontologies and hierarchically organised controlled vocabularies is fundamental to help the subject matter experts to find the right information," said Jeff Spitzner, President of Rescentris Inc. "Cloud computing and semantic web technologies are now becoming mainstream in leading research organisations and we have been applying these over the past decade to turn data into shared and actionable knowledge."

Semantic technologies are gaining momentum. Controlled vocabularies, controlled protocols and ontologies will enable scientists, students and other experts to perform scientific knowledge modelling, and logic-based hypothesis checking with significantly better results. Even simple terms like 'gene' and 'cancer' mean different things for different kinds of research activities and data types. Use of namespaces helps distinguish between universal properties like pH or temperature, properties that mean different things in different spaces, such as 'sequence', which could be DNA bases or the order of treatments, and properties that are specific to one context like 'activity', which could refer to the results measured in a specific assay. When cloud-based data is semantically annotated it can be linked back to any other scientific lab data system and provenance (the evidence trail) can be tracked.

The impact of semantic web technology in the cloud is significant. Semantic web technology will considerably decrease the need for central warehouses of information. There will be no need to transfer large amounts of data across networks. This truly enhances collaboration as semantic agents created by scientists would constantly lurk the internal and external web for information related to a molecule or experiment. Users would be instantly alerted to new data sources and have it automatically categorised, prioritised and sorted based on the rules set by the scientists.

"Service is as important as the product itself," said Megean Schoenberg, Director of Enterprise Software at PerkinElmer. "Scientific research will be accelerated when there is a close contact between scientists, user community and software solution providers." PerkinElmer recently acquired CambridgeSoft, ArtusLabs and Labtronics.

Building bridges

New initiatives between industry, solution providers and government are developing which show commitment to implement the cloud to support cross-functional collaborations. IDBS, a global provider of data management, analytics and modelling solutions, announced this summer that it is leading a stratified medicine consortium to build a UK-wide Cancer Research and Collaboration Platform. This highly secured data solution brings together patient, 'omic and up-to-date research data to enable cohort selection, scientifically-based data mining and the secure sharing of cohort outcomes. "The Cloud provides that extensible, accessible computing capability which, with strong application security, can provide the necessary Infrastructure-as-a-Service and Platform-as-a-Service systems which will be a feature of tomorrow's software landscape," said Chris Molloy, VP Corporate Development, IDBS. It helps clinical-researchers better identify patients by their clinical and genetic profiles and examine the effects of genetics on disease and on disease outcomes – fundamentally a collaborative effort between clinicians.

Hybrid clouds take advantage of the scalability and cost-effectiveness of a public cloud without exposing mission-critical applications and data to third-party vulnerabilities. These cloud offerings will allow individuals to link computing resources across their organisations into a single, high-performance, private cloud while providing system administrators the flexibility to set priorities based on business or technical needs. IBM recently announced a High Performance Cloud (HPC) offering that enables clients to manage and prioritise HPC assets on a global basis while maintaining operations and data securely behind its firewall.

Eureka moments

The cloud stimulates agile collaboration. Instead of disappearing into a black-hole for 18 months and not showing up with any results until the first project milestone, the remote teams are fully real-time

Informatics

Reference
1 Graham, M. Engineering Earth, 2011. geospace.co.uk.

integrated into the project team. Instead of just getting n-point results in a spreadsheet that was emailed to you in the morning, the cloud enables you to see the details of how the experiment was set up, what the observations were as it was being run with off-site collaborators, at any time at any place. It is in the same format, the same template and has the same business rules attached to it. No need to go through a procedure of mental mapping of what that data means. It integrates the external research teams with your own organisation by having access to information on a real-time basis.

Social media such as Facebook, a great public cloud example, is starting to become an extension of reaching people who are located in remote places. Neglected diseases are an area where more community come together helped by The Gates Foundation and from companies including Pfizer, Lilly, Merck and GSK.

Full circle
In 2009, human beings generated more data than all previous 5,000 years combined. Data Intensive Science is becoming mainstream. Technology will change the dynamics of how scientists work together. The cloud is not just an IT initiative, it really changes the way people and science can work together and how they can collaborate globally, in real-time. No need to wait for months for the scientific paper to be published. Building trust relationships to make these teams communicate with each another still remains a people issue. That's how thousands of years ago experimental science was invented. Some things will never change...   DDW

Peter Boogaard is the founder of Industrial Lab Automation, an independent consulting firm that provides management services to address harmonisation, integration and consolidation of business processes in Life Science development and manufacturing to enable cross-functional collaboration between research, development, quality assurance and manufacturing corporations to achieve Quality by Design (QbD) initiatives. Email: peter-boogaard@IndustrialLabAutomation.com. www.iLabAutomation.com. www.industriallabautomation.com/cloud.php