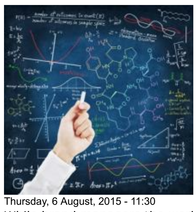
This is what Big Data really looks like: CERN, the universe and everything



Thursday, 6 August, 2015 - 11:30

While brand managers the world over complain about the deluge of data they need to make sense of these days, data scientists at CERN are trying to solve the mysteries of the universe using facilities like the Large Hadron Collider (LHC), the world's largest particle accelerator. Sifting through billions of data points from a fire hose measurable in terabytes per second, the data challenges faced by CERN's physicists dwarf those of most commercial entities.

Which-50 interviewed Bob Jones, Project Leader at CERN, who is a driving force behind CERN's information management expertise and was the Head of CERN openlab between January 2012 and December 2014.

Among the many challenges Jones and his colleagues face is trying to gather insights from more than 20 petabytes of data from CERN's Large Hadron Collider every year and, in particular, isolating the small number of particle collisions involving the elusive Higgs particle from the vast stream of event data.

To manage this scale of data effectively, CERN has been a long-time champion of distributed processing and innovative data storage approaches.

Jones works in the IT department of CERN (http://information-technology.web.cern.ch/about), which serves the whole organisation and has a critical role to play in supporting the scientific programme.

Jones described what for him and his colleagues are the biggest challenges of big data at CERN. The challenges are many and varied: "It is the combination of storage capacity, access patterns and sometimes unpredictable analysis workloads that are the biggest challenge," he said.

As well as dealing with the voluminous data produced by CERN's many experiments, the speed with which physicists develop and change focus in their experimental work adds to the data-management challenge. According to Jones, research moves quickly and those involved can't always predict which particular dataset will prove to be the most popular and require the most resources.

"We are ready for Run 2 of the LHC (http://home.web.cern.ch/about/updates/2015/06/lhc-experiments-back-business-record-energy), which started in June 2015 with the experiments taking data at the unprecedented energy of 13 TeV, following the two-year long shutdown. We have multiple 100Gbps lines linking the two centres, which enables us to operate them as a single OpenStack cloud (https://www.openstack.org/user-stories/cern/)."

And the volume of data is set to continue growing. The LHC experiments have already recorded 100 times more data for the summer conferences this year than they had around the same time after the LHC started up at 7 TeV in 2010, he says.

On the grid

While CERN originally hosted all IT services on-site in a traditional service provision model, as the needs of the scientific programme expanded it focused more heavily on developing off-site data processing and management capability through grid computing.

This approach has allowed CERN to federate IT resources from partner organisations around the world, as well as scale storage and processing power more efficiently. The cloud services market — originated by Amazon Web Services back in 2002 and much-loved by data scientists in publishing and other high-data-volume sectors — has proved a powerful tool for CERN as well.

"This market now offers us new opportunities to increase the scale and range of the IT services we can build on. We are working towards a hybrid cloud model, where we can we can opportunistically use any resources taking into account availability, price and policy."

"The volume of data produced at the LHC is a challenge, but the process is similar for smaller-scale experiments," he said

The raw data per event is around one million bytes (1 MB), produced at a rate of about 600 million events per second. The Worldwide LHC Computing Grid tackles this mountain of data in a two-stage process.

First, it runs dedicated algorithms written by physicists to reduce the number of events and select those considered interesting — a sophisticated winnowing out of noise from the data sets. "Analysis can then focus on the most important data — that which could bring new physics measurements or discoveries."

When it comes to analytical tools, it's unsurprising to hear that CERN's data scientists have built and tweaked their own analytical toolkit. "The physics community has progressively developed, over a number of years, a set of software tools dedicated to this task. These tools are constantly being improved to ensure they continue work at the growing scale of the LHC data challenge. ROOT (https://root.cern.ch/drupal/) is a popular data-analysis framework — it is a bit like R on steroids."

Jones says that grid computing has also been immensely helpful in enabling physicists to run analysis at scale. "Grid computing helps by providing an underlying global infrastructure with the capacity to be able to match the analysis needs of the LHC. But the grid itself is evolving to make more use of cloud computing techniques and profit from the improvements in hardware (processors, storage etc.) as well as the cost-effectiveness of high performance networks."

Lessons for brands

The CERN Data Centre has the ability to process incredibly high throughput in order to manage the data coming out of the Large Hadron Collider. That prompts the question of whether there will be many situations in which the commercial sector would need that extreme throughput capability.

According to Jones, "CERN is a leader but not alone in having to deal with such high data throughputs. We expect to see similar scales in other sciences (such as next generation genome sequencing (http://www.nature.com/nature/journal/v498/n7453/full/498255a.html) as well as the Square Kilometre Array (https://www.skatelescope.org/) which will primarily be deployed in Australia and South Africa) and various business sectors linked to the growing Internet of Things (http://www.forbes.com/sites/jacobmorgan/2014/05/13/simple-explanation-internet-things-that-anyone-can-understand/) in the near future."

He described CERN as being ahead of the curve, and said the technologies and processes developed — as well as the lessons learned — at CERN can be applied in other fields. However, he emphasised that CERN's advanced capabilities are not acquired by happenstance — the organisation spends a great deal of effort in growing the skills needed to develop cutting-edge data solutions.

"Education is a key element of CERN's mission. For those working at CERN, we have technical and management training programmes (http://hr-dep.web.cern.ch/LD/learning-development) and series of computing seminars as well as the CERN School of Computing (https://csc.web.cern.ch/). We are constantly recruiting young scientists, engineers and technicians who also bring new skills and ideas into CERN's environment. Engagement with leading IT companies through CERN openlab (http://openlab.web.cern.ch/) has been a source of many new developments and helps train successive generations of personnel in the latest techniques," he said.

SMEs

- User Stories (/taxonomy/term/85)
- European CloudScout (/cloudscout)
- Managing Risk (/business-guide-managing-risk)
- SME Cloud Security Tool (http://www.enisa.europa.eu/activities/Resilience-and-CIIP/cloud-computing/security-for-smes/sme-guide-tools)
- Legal services (/legal-services)
- Cloud Standards Guide (/cloud-standards-guide)

Governments

- User Stories (/gov-user-stories)
- Cloud Procurement Guides (/procurement)
- Managing Risk (/business-guide-managing-risk)
- PICSE Wizard - Cloud Procurement Made Easy (http://wiz.picse.eu/#/app/home?lang=en&code=en)
- Managing Risk (/gov-risk-management)
- Cloud security certifications (https://resilience.enisa.europa.eu/cloud-computing-certification/list-of-cloud-certification-schemes/cloud-certification-schemes-metaframework)

CloudWATCH for ...

- Cloud Trends (/trends)
- Guide to Cloud Certifications (/certifications)
- Legal Tips for the Cloud (/legal-tips)
- Interoperable Clouds (/interoperability)
- Guide to Cloud Standards (/standards-0)
- Standards Testing (/standards-testing)

News

- Azure Webinar: Big-Data Research with Hadoop on Azure - Sept 22 (/azure-webinar-big-data-research-hadoop-azure-sept-22)
- Join to discussion at NIST Cloud Computing Forum & Workshop (/join-discussion-nist-cloud-computing-forum-workshop)
- CALL FOR PAPERS at CLASS conference 2016, 8-9 November 2016, Ljubljana (/call-papers-class-conference-2016-8-9-november-2016-ljubljana)
- Estonia chooses Ericsson to lead public sector cloud transformation (/estonia-chooses-ericsson-lead-public-sector-cloud-transformation)
- CloudWATCH2 & SNIA Cloud Interoperability Plugfest 24 Sept. 19-21 2016 (/cloud-interoperability-plugfest-24-sept-19-21-2016)

Events

- CloudWATCH2 Cloud Security Plugfest – Madrid 14 Nov 2016 (/cloudwatch2-cloud-security-plugfest-%E2%80%93-madrid-14-nov-2016)

Read More (/our-events)

To read the full story, please visit Which-50 (http://which-50.com/blog/2015/august/03/this-is-what-big-data-really-looks-like-cern-the-universe-and-everything/).

Image courtesy of MR LIGHTMAN at FreeDigitalPhotos.net

Good to know (/side-menu/good-know)
Category:
Trends (/trends)

Copyright 2016 CloudWATCH2 has received funding from the European Union's Horizon 2020 programme - DG CONNECT Software & Services, Cloud.
Contract No. 644748