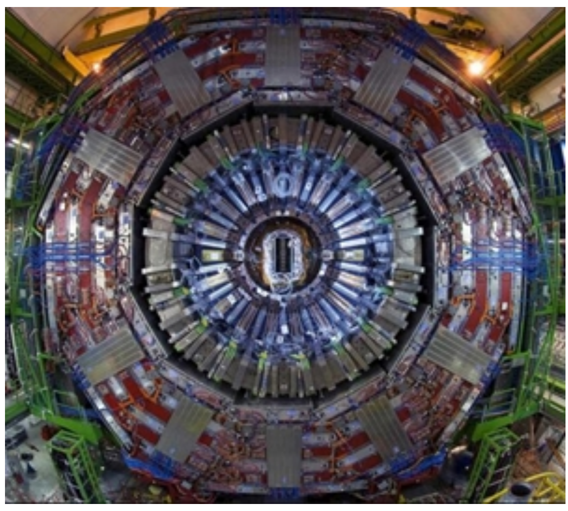
This is what Big Data really looks like: CERN, the universe and everything



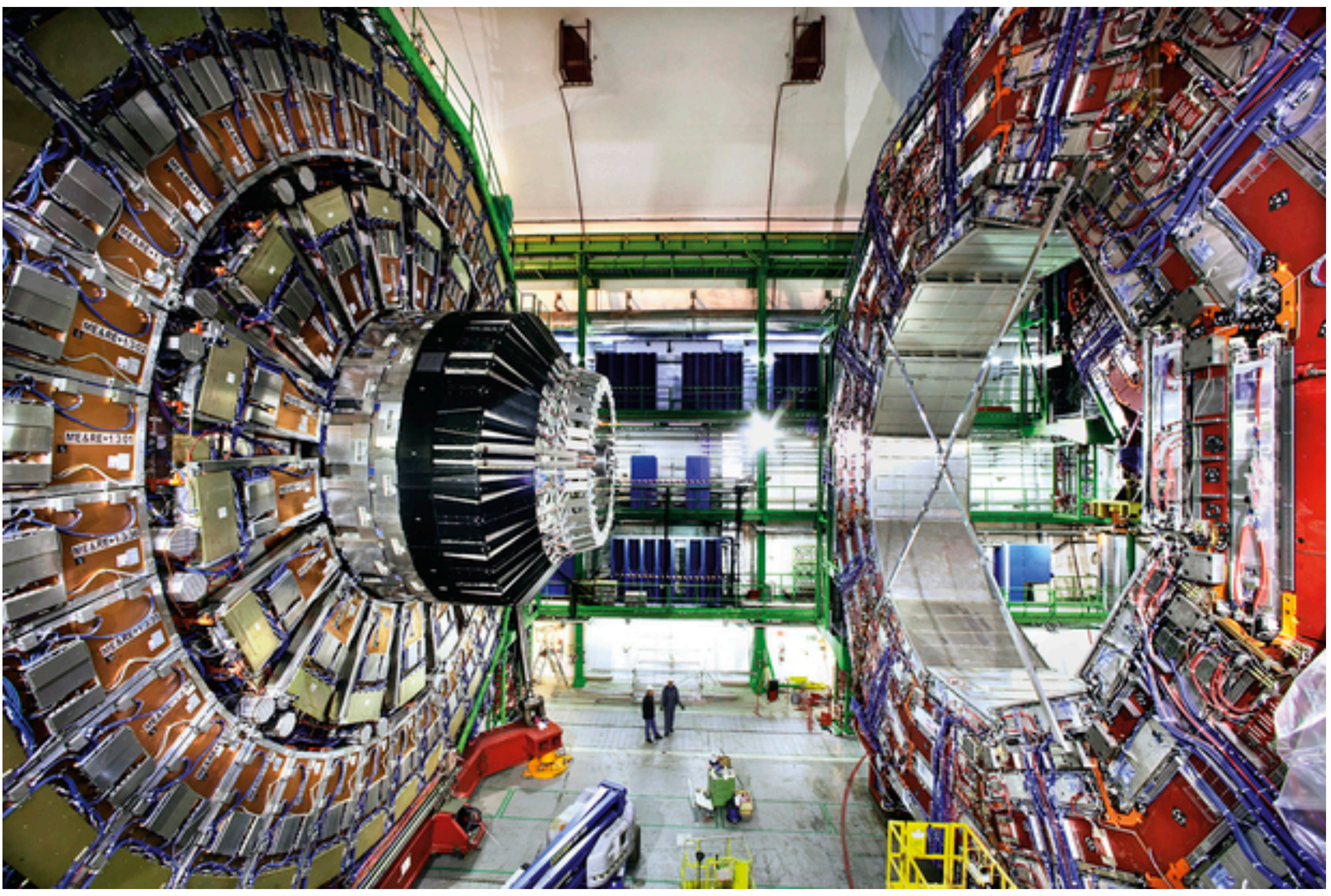
BY ANDREW BIRMINGHAM AND ANNA RUSSELL, MONDAY AUGUST 3, 2015

So you think you know all about analytics? Patting yourself on the back about your latest cross-channel attribution modeling and the terabytes of data you've successfully corralled into a database? Time for a little perspective — because there's big data and then there's BIG data.

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 quantify the magnitude of this task, Jones explains: "*During Run 1, the LHC produced six million billion proton-proton collisions… Of these, only around 400 produced results compatible with the Higgs particle, whose discovery was announced in July 2012. So you can see that identifying the right 400 events out of six million billion proton-proton collisions is really like looking for a needle in a haystack.*"

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

According to Jones, "*CERN and the physics community has been a driving force in the development of grid computing since the year 2001. This has led to the deployment of a production infrastructure with a global footprint known as the Worldwide LHC Computing Grid, or WLCG for short, which provides the resources to store, distribute and analyse the data from the LHC.*"

Jones works in the IT department of CERN, which serves the whole organisation and has a critical role to play in supporting the scientific programme. Understandably, "*It is a very demanding environment with continuous renewal and upgrades to services.*"

Despite — or perhaps because of — this pressure, Jones says he is constantly impressed by the quality of the people he works with, including the numerous world-class experts on site, and their approach to tackling vast and complex tasks. "*CERN has a university campus feel about it and the people are very open and willing to help and collaborate.*"

We asked Jones to describe 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.

"*So our system needs to be very dynamic. We have a 3.5 MW computer centre on-site in Geneva and have leased space in a second computer centre at Wigner in Budapest, Hungary.*

"*We are ready for Run 2 of the LHC, 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.*"

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.*"

He said CERN is actively investigating this approach with cloud service companies and other research organisations in the context of the Helix Nebula initiative. "*So seeking new opportunities and keeping flexible enough to profit from them is a key aspect of our strategy at CERN. But we have learnt that operating production services at this scale is not something that can be improvised.*"

According to Jones, "*Every time we have had to increase scale it has required development which takes time and advanced planning. Similarly, the importance of preserving data is paramount. CERN puts significant resources into bit-level preservation of data, including the use of tape systems where the technology continues to evolve.*"

This archived data must be actively managed to ensure it remains available for future use, he said. However there is most certainly a balancing act required to ensure CERN holds on to meaningful scientific data but doesn't unintentionally house excessive volumes of meaningless information.

"*At this scale it is not possible to keep all the data (the LHC produces up to a Petabyte of data per second) and it is essential to have efficient data-filtering mechanisms so that we can separate the wheat from the chaff. A key risk is throwing away the data you need and cannot reproduce.*"

We also asked Jones if the approach to data analysis or exploration is very different when dealing with the vast quantities of data from the Large Hadron Collider or if the thought processes are similar to smaller-scale experiments, just executed using tools capable of handling greater scale.

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

"*Particles collide at high energies inside the detectors, creating new particles that decay in complex ways as they move through layers of subdetectors. The subdetectors register each particle's passage and microprocessors convert the particles' paths and energies into electrical signals, combining the information to create a digital summary of the collision event.*"

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 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 as well as the Square Kilometre Array which will primarily be deployed in Australia and South Africa) and various business sectors linked to the growing Internet of Things 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 and series of computing seminars as well as the CERN School of Computing. 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 has been a source of many new developments and helps train successive generations of personnel in the latest techniques,* " he said.

CERN is also a poster child for the power of not only open source but also a culture of organisational openness. Jones said, "*CERN's open culture coupled with developments such as commercial cloud services where an organisation's data may be stored off-site, and a Bring Your Own Device (BYOD) policy for the site, means we have to be proactive to ensure everyone respects intellectual property rights and the relevant data protection legislation.*

"*We are also active in the deployment of federated identity-management systems for access to IT services, and such a model has been in place for the Worldwide LHC Computing Grid since its creation.*"

Bob Jones is speaking at 9.15am at the IAPA National Conference/ADMA Advancing Analytics event at the Sydney Hilton Hotel tomorrow.

About the authors

Andrew Birmingham is the director of the Which-50 Digital Intelligence Unit. Anna Russell is a Sydney-based data scientist and the director of Polynomial

WHICH-50
DIGITAL
INTELLIGENCE
UNIT

ADMA is a corporate member of the Which-50 Digital Intelligence Unit. Members contribute their expertise and insights to Which-50 for the benefit of our senior executive audience. Membership fees apply.

BY ANDREW BIRMINGHAM AND ANNA RUSSELL, FOLLOW @WHICH50 FOR MORE GREAT INSIGHTS

TAGS

- BIG DATA
- CERN

11 +11

SIGN UP TO OUR IRREGULAR INSIGHTS NEWSLETTER

First Name

Last Name

Email Address

Job Title

SIGN UP

RELATED STORIES MOST POPULAR

- The measurement dilemma – it's not digital that's the problem. Actually, it's you
- McKinsey: Big Data could unlock $200B in marketing value
- If your instincts tell you to trust the data, listen. If not, then don't trust your instincts
- Can you make data sexy? Yes, if you treat it right
- Data must accord with managing customer preferences