2
CERN

Unravelling The Secrets Of The Universe With Big Data

Background

CERN are the international scientific research organization that operate the Large Hadron Collider (LHC), humanity's biggest and most advanced physics experiment. The colliders, encased in 17 miles of tunnels buried 600 feet below the surface of Switzerland and France, aim to simulate conditions in the universe milliseconds following the Big Bang. This allows physicists to search for elusive theoretical particles, such as the Higgs boson, which could give us unprecedented insight into the composition of the universe.

CERN's projects, such as the LHC, would not be possible if it weren't for the Internet and Big Data – in fact, the Internet was originally created at CERN in the 1990s. Tim Berners-Lee, the man often referred to as the "father of the Internet", developed the hypertext protocol which holds together the World Wide Web while at CERN. Its original purpose was to facilitate communication between researchers around the globe.

The LHC alone generates around 30 petabytes of information per year – 15 trillion pages of printed text, enough to fill 600 million filling cabinets – clearly Big Data by anyone's standards!

In 2013, CERN announced that the Higgs boson had been found. Many scientists have taken this as proof that the standard model of particle physics is correct. This confirms that much of what we think we know about the workings of the universe on a subatomic level is essentially right, although there are still many mysteries remaining, particularly involving gravity and dark matter.

What Problem Is Big Data Helping To Solve?

The collisions monitored in the LHC happen very quickly, and the resulting subatomic "debris" containing the elusive, sought-after particles exists for only a few millionths of a second before they decay. The exact conditions that cause the release of the particles which CERN are looking for only occur under very precise conditions, and as a result many hundreds of millions of collisions have to be monitored and recorded every second in the hope that the sensors will pick them up.

The LHC's sensors record hundreds of millions of collisions between particles, some of which achieve speeds of just a fraction under the speed of light as they are accelerated around the collider. This generates a massive amount of data and requires very sensitive and precise equipment to measure and record the results.

How Is Big Data Used In Practice?

The LHC is used in four main experiments, involving around 8000 analysts across the globe. They use the data to search for elusive theoretical particles and probe for the answers to questions involving antimatter, dark matter and extra dimensions in time and space.

Data is collected by sensors inside the collider that monitor hundreds of millions of particle collisions every second. The sensors pick up light, so they are essentially cameras, with a 100-megapixel resolution capable of capturing images at incredibly high speeds.

This data is then analysed by algorithms that are tuned to pick up the telltale energy signatures left behind by the appearance and disappearance of the exotic particles CERN are searching for.

The algorithms compare the resulting images with theoretical data explaining how we believe the target particles, such as the Higgs boson, will act. If the results match, it is evidence the sensors have found the target particles.

What Were The Results?

In 2013, CERN scientists announced that they believed they had observed and recorded the existence of the Higgs boson. This was a huge leap forward for science as the existence of the particle had been theorized for decades but could not be proven until technology was developed on this scale.

The discovery has given scientists unprecedented insight into the fundamental structure of the universe and the complex relationships between the fundamental particles that everything we see, experience and interact with is built from.

Apart from the LHC, CERN has existed since the 1950s and has been responsible for a great many scientific breakthroughs with earlier experiments, and many world-leading scientists have made their name through their work with the organization.

What Data Was Used?

Primarily, the LHC gathers data using light sensors to record the collision, and fallout, from protons accelerated to 99.9% of the speed of light. Sensors inside the colliders pick up light energy emitted during the collisions and from the decay of the resulting particles, and convert it into data which can be analysed by computer algorithms.

Much of this data, being essentially photographs, is unstructured. Algorithms transform light patterns recorded by the sensors into mathematical data. Theoretical data – ideas about how we think the particles being hunted will act – is matched against the sensor data to determine what has been captured on camera.

What Are The Technical Details?

The Worldwide LHC Computing Grid is the world's largest distributed computing network, spanning 170 computing centres in 35 different countries. To develop distributed systems capable of analysing 30 petabytes of information per year, CERN instigated the openlab project, in collaboration with data experts at companies including Oracle, Intel and Siemens. The network consists of over 200,000 cores and 15 petabytes of disk space.

The 300 gigabytes per second of data provided by the seven CERN sensors is eventually whittled down to 300 megabytes per second of "useful" data, which constitutes the product's raw output. This data is made available as a real-time stream to academic institutions partnered with CERN.

CERN have developed methods of adding extra computing power on the fly to increase the processing output of the grid without taking it offline, in times of spikes in demand for computational power.

Any Challenges That Had To Be Overcome?

The LHC gathers incredibly vast amounts of data, very quickly. No organization on earth has the computing power and resources necessary to analyse that data in a timely fashion. To deal with this, CERN turned to distributed computing.

They had already been using distributed computed for some time. In fact, the Internet as we know it today was initially built to save

scientists from having to travel to Geneva whenever they wanted to analyse results of CERN's earlier experiments.

For the LHC, CERN created the LHC Distributed Computing Grid, which comprises 170 computer centres in 35 countries. Many of these are private computing centres operated by the academic and commercial organizations partnered with CERN.

This parallel, distributed use of computer processing power means far more calculations per second can be carried out than even the world's most powerful supercomputers could manage alone.

What Are The Key Learning Points And Takeaways?

The groundbreaking work carried out by CERN, which has greatly improved our knowledge of how the universe works, would not be possible without Big Data and analytics.

CERN and Big Data have evolved together: CERN was one of the primary catalysts in the development of the Internet which brought about the Big Data age we live in today.

Distributed computing makes it possible to carry out tasks that are far beyond the capabilities of any one organization to complete alone.

REFERENCES AND FURTHER READING

Purcell, A. (2013) CERN on preparing for tomorrow's big data, http://home.web.cern.ch/about/updates/2013/10/preparing-tomorrows-big-data

Darrow, B. (2013) Attacking CERN's big data problem, https://gigaom.com/2013/09/18/attacking-cerns-big-data-problem/

O'Luanaigh, C. (2013) Exploration on the big data frontier, http://home.web.cern.ch/students-educators/updates/2013/05/exploration-big-data-frontier

Smith, T. (2015) Video on CERN's big data, https://www.youtube.com/watch?v=j-0cUmUyb-Y