Big Data Analysis using Cloud Technology

SCIFM0004 - Distributed Computing Mini-project

Introduction

In this mini project you will analyse data from a particle physics experiment using cloud technology. The goal is to construct an application which can (in principle) be scaled from your laptop to a large cluster. You are provided with datafiles and example analysis application, as described below.

The ATLAS Experiment

Data handling is a significant fundamental challenge associated with the Large Hadron Collider, as well as other 'big science' experiments such as the Square Kilometer Array. In order to detect rare processes, the LHC needs to collide protons at a rate of 10^9 Hz over many years. And in order to extract useful information from those collisions, they must be examined with very high precision detectors. The resulting data volumes are extraordinary, and dedicated computing techniques have been developed to achieve this.

ATLAS is an experiment at the Large Hadron Collider. It can be thought of as a complex 3-dimensional digital camera, comprising $\mathcal{O}$(100M) channels, producing data at an effective rate of $\mathcal{O}$(1PB/s). This is far in excess of what can be stored and analysed, so a pseudo real-time analysis of each event is performed by a so-called "trigger" system, in a data center next to the experiment, which selects events of interest and reduces the data rate to a more manageable 10TB/day. The annual dataset from CMS is up to 20PB, which is stored, processed and analysed in a large number of data centers around the world, using cloud technology.

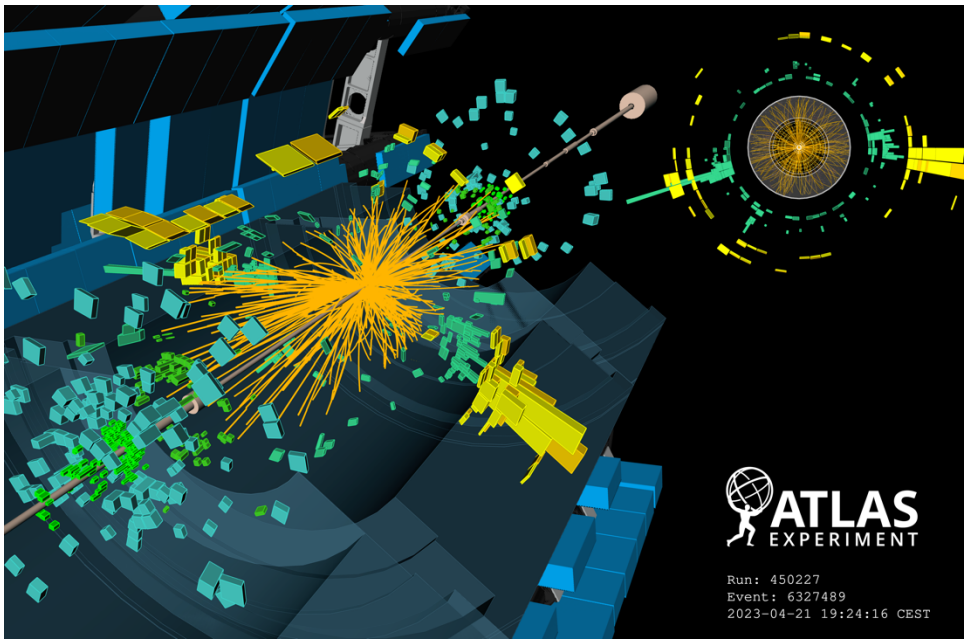

Figure 1 - An ATLAS collision event

Data Processing

You will implement a data processing task using publicly available data from the ATLAS experiment [1]. The task includes processing data stored in a large number of files, selecting events of interest, and producing summary information (histograms).

The ATLAS Open Data includes 12 example physics analyses, which are described in [2] with corresponding example notebooks in [3]. The datasets themselves can be found at [4].

An example notebook which performs the analysis can be found at : https://github.com/atlas-outreach-data-tools/notebooks-collection-opendata/blob/master/13-TeV-examples/uproot_python/HZZAnalysis.ipynb

It is worth noting that the ATLAS Open Datasets uses greatly simplified data content and formats compared with the datasets used by the collaboration. In particular, the Open Data is the result of very substantial amounts of processing, including reconstruction and calibration. The analysis examples given represent the very final step in the data processing chain. However, the techniques you will use in this exercise are also used to manage real LHC data processing workflows across the LHC Computing Grid, which comprises 1.4M cores and 1.5 Exabytes of storage worldwide.

Aims of the Project

The goal of the project is to develop a processing system which can analyse ATLAS data using the cloud tools introduced in the unit. The system should :

1. Configure itself automatically, with minimum human intervention.
2. Perform the analysis processing in a distributed fashion, using multiple (virtual) nodes.
3. Be scalable (in principle) to a much larger system capable of processing much larger data volumes.

You are not expected to *test* the scalability of your system, since we are unable to provide you with a data center. However, you should consider the potential for bottlenecks and constraints in your system, and how you avoid them when scaling up to a larger number of nodes, and discuss these issues in your report.

Getting Started

The first steps of your project could include the following.

1. Select an analysis, download the corresponding example notebook, and verify you can run it.
2. Understand the steps involved in the analysis and decide how you will break it down into individual tasks, and whether they must be run in parallel, or in sequence.
3. Think about how you will map these tasks onto processing units.
4. Decide how you will use cloud technology to launch, configure and control your processing units.

You can use any of the software packages introduced in the unit. If you have questions about whether other packages are acceptable, please discuss with your instructor.

Submission

You should submit your work via the Blackboard submission point. Please submit both

1. An archive of your code repository. This should include all files necessary to create and run your distributed analysis system, together with a readme file which explains how to run the code.
2. A report in PDF format. Your report should not exceed 2000 words.

References

[1] ATLAS collaboration, *Review of the 13 TeV ATLAS Open Data release*, ATL-OREACH-PUB-2020-001, https://cds.cern.ch/record/2707171

[2] ATLAS collaboration, *13 TeV ATLAS Open Data physics analysis examples,* https://opendata.atlas.cern/docs/physics/intro

[3] ATLAS collaboration, https://github.com/atlas-outreach-data-tools/notebooks-collection-opendata/tree/master/13-TeV-examples/uproot_python

[4] ATLAS collaboration, *ATLAS Open Datasets in ROOT format,* https://opendata.atlas.cern/docs/datasets/files