**ABSTRACT**

Public clouds have democratised the access to analytics for virtually any

institution in the world. Virtual Machines (VMs) can be provisioned on demand, and be used to crunch data after uploading into the VMs. While this task is trivial for a few tens of VMs, it becomes increasingly complex and time consuming when the scale grows to hundreds or thousands of VMs crunching tensor hundreds of TB. Moreover, the elapsed time comes at a price: the cost of provisioning VMs in the cloud and keeping them waiting to load the data. In this project we present a big data provisioning service that incorporates hierarchical and peer-to-peer data distribution techniques to speed-up data loading into the VMs used for data processing. The system dynamically mutates the sources of the data for the VMs to speed-up data loading. We tested this solution with 1000 VMs and 100 TB of data, reducing time by at least 30 % over current state of the art techniques. This dynamic topology mechanism is tightly coupled with classic declarative machine configuration techniques (the system takes a single high-level declarative configuration file and configures both software and data loading). Together, these two techniques simplify the deployment of big data in the cloud for end users who may not be experts in infrastructure management.