Otwarta nauka w chmurze: w kierunku uniwersalnej platformy obliczeń naukowych i statystycznych

Wielka Brytania za pośrednictwem programu e-Science, USA za pośrednictwem infrastruktury cybernetycznej finansowanej przez NSF, a Unia Europejska za pośrednictwem ICT Calls mających na celu zapewnienie „technologicznego rozwiązania problemu wydajnego łączenia danych, komputerów i ludzi w celu umożliwienie wyprowadzania nowych teorii naukowych i wiedzy”1. Sieć Grid, przewidziana jako główny akcelerator odkryć, nie spełniła oczekiwań, które wzbudzała na początku i nie została przyjęta przez szeroką populację naukowców. Siatka jest dobrym narzędziem dla fizyków cząstek elementarnych i pozwoliła im stawić czoła ogromnym wyzwaniom obliczeniowym związanym z ich dziedziną. Jednak jako technologia i paradygmat dostarczania przetwarzania na żądanie nie działa i nie można go naprawić. Z jednej strony „abstrakcje, które udostępniają Grids - użytkownikowi końcowemu, wdrażającym i twórcom aplikacji – są nieodpowiednie i muszą być na wyższym poziomie”, a z drugiej strony, sieci akademickie są z natury niezrównoważone ekonomicznie. Nie mogą konkurować z usługą zleconą na zewnątrz Przemysłowi, której jakość i cena byłyby napędzane przez siły rynkowe. Technologie wirtualizacji i ich następstwo, chmura typu Infrastructure-as-a-Service (IaaS), obiecują umożliwić to, czego nie udało się osiągnąć Grid: zrównoważone środowisko dla nauk obliczeniowych, które obniżyłoby bariery dostępu do sfederowanych zasobów obliczeniowych, narzędzia i dane oprogramowania; umożliwiają współpracę i udostępnianie zasobów oraz dostarczają elementy budulcowe wszechobecnej platformy do identyfikowalnych i odtwarzalnych badań obliczeniowych. Amazon Elastic Compute Cloud (EC2) to przykład infrastruktury jako usługi, z którego może korzystać każdy dzisiaj. Jej znaczny sukces zapowiada nadejście nowej ery. Jednak wprowadzenie tej ery dla badań i edukacji nadal wymaga oprogramowania, które wypełniłoby lukę między chmurą a codziennymi narzędziami naukowców i uczyniłoby z infrastruktury jako usługi towar trywialny . W tym artykule opisano platformę Elastic-R,3,4, która sprawia, że praca z R w chmurze jest tak prosta, jak praca z nim lokalnie. Mówiąc ogólniej, ma być brakującym ogniwem między chmurą a najczęściej używanymi narzędziami do analizy danych i naukowymi środowiskami obliczeniowymi (SCE). Elastic-R synergizuje scenariusze użytkowania tych środowisk ze scenariuszami użytkowania chmury i zapewnia im to, co chmura ma najlepiej do zaoferowania Przyjazny dla użytkownika i elastyczny dostęp do infrastruktury jako usługi: interfejsy chmury są proste i eksponowane właściwe abstrakcje do zarządzania i korzystania z urządzeń wirtualnych w sfederowanym środowisku obliczeniowym, ale konsole w chmurze pozostają narzędziami dla znawców komputerów. Elastic-R oferuje uproszczoną fasadę chmury, dzięki której każdy naukowiec może wybrać i uruchomić maszynę wirtualną ze specyficznym naukowym środowiskiem obliczeniowym (przykład: R wersja 2.9, Scilab, Sage, Root itp.). Naukowiec może wtedy mieć dostęp do pełnych możliwości środowiska za pomocą środowiska pracy Elastic-R Java lub z poziomu standardowej przeglądarki internetowej za pomocą środowiska pracy Elastic-R Ajax. Naukowiec może wydawać polecenia, instalować i używać nowych pakietów, generować i wchodzić w interakcję z grafiką, przesyłać i przetwarzać pliki, pobierać wyniki, tworzyć i edytować arkusze kalkulacyjne po stronie serwera z obsługą R itp. Naukowiec może odłączyć się od silnika i ponownie połączyć się z w dowolnym miejscu, pobiera pełną sesję, w tym obszar roboczy, grafikę itp., i może kontynuować pracę od miejsca, w którym przerwał. Maszynę wirtualną można po prostu wyłączyć, gdy nie jest już potrzebna. Użytkownik jest obciążany tylko za czas użytkowania. Dane użytkownika (zawartość katalogu roboczego) pozostają na dysku wirtualnym w chmurze, który można ponownie podłączyć do nowej instancji maszyny wirtualnej.

Współpraca: instancja maszyny wirtualnej w chmurze ma publiczny adres IP i może być widoczna i używana przez współpracowników właściciela, którzy mogą znajdować się w dowolnym miejscu. Elastic-R pozwala naukowcowi na udostępnienie swojej maszyny i sesji R (na przykład) współpracownikom. Każdy z nich może podłączyć swoje środowiska pracy Java lub Ajax do tego samego silnika R i kontrolować ten silnik oraz aktualizować jego środowisko. Działania każdego

współpracownika w konsoli (wydawane polecenia, czat), na grafice (drukowanie, adnotacje, zmiana rozmiaru i przeglądanie slajdów) lub w arkuszach kalkulacyjnych (aktualizacja komórek, zaznaczanie komórek) są transmitowane do innych, a ich stanowiska pracy pokazują zmiany w czasie rzeczywistym.

Elastyczność na żądanie: Elastic-R udostępnia naukowcom ważną cechę chmury, którą jest możliwość wyboru pojemności instancji maszyn wirtualnych, takich jak liczba wirtualnych rdzeni, rozmiar pamięci i miejsce na dysku. Następnie może przeprowadzać analizy lub symulacje w chmurze, które wymagają więcej pamięci niż jest dostępne na jego laptopie lub zajęłyby dni, gdyby były uruchamiane lokalnie. Elastic-R umożliwia naukowcom rozwiązywanie problemów wymagających dużej mocy obliczeniowej poprzez uruchamianie dowolnej liczby maszyn wirtualnych z silnikami R, które mogą przetwarzać równolegle zadania częściowe. Te pule silników mogą być również używane do tworzenia aplikacji internetowych z dynamiczną zawartością analityczną generowaną przez R lub dowolne inne środowisko. W przypadku tych aplikacji platforma Elastic-R umożliwia cloudbursting: maszyny wirtualne mogą być uruchamiane lub wyłączane (zwiększając lub zmniejszając rozmiar puli silników) w celu skalowania w górę lub w dół w zależności od obciążenia aplikacji.

Elastyczność wdrażania aplikacji: Chmura może bardzo łatwo obsługiwać aplikacje typu klient-serwer. Elastic-R to platforma, która pozwala każdemu na gromadzenie metod statystycznych/numerycznych i danych na serwerze (instancja maszyny wirtualnej Elastic-R w chmurze) oraz wizualne tworzenie i publikowanie, w postaci adresów URL, interaktywnych interfejsów użytkownika i pulpitów nawigacyjnych ujawniając te metody i dane. Elastic-R zapewnia również narzędzia, które pozwalają każdemu ujawnić te metody (na przykład zaimplementowane przez funkcje R) jako usługi sieci Web SOAP, które mogą być używane jako usługi obliczeniowe w chmurze dla potoków analizy danych lub jako węzły dla środowisk roboczych przepływu pracy.

Możliwości nagrywania: Ponieważ naukowe środowiska obliczeniowe dostępne za pośrednictwem Elastic-R działają na maszynach wirtualnych, a katalogi robocze są hostowane na dyskach wirtualnych, w dowolnym momencie można utworzyć migawkę w pełni zaktualizowanego środowiska obliczeniowego. Ta migawka może zostać zarchiwizowana lub udostępniane każdemu korzystającemu ze stołów warsztatowych Elastic-R: autor może dzielić się swoim środowiskiem z recenzentami czasopisma, do którego przesłał swoją pracę, nauczyciel może udostępnić swoim studentom środowisko nauki statystycznej potrzebne do jego kursu, badacz w laboratorium A może udostępnić swoje środowisko symulacyjne współpracownikom w laboratorium B itp.

Otwarta platforma obliczeń naukowych, bloki konstrukcyjne

R to język i środowisko do obliczeń statystycznych i grafiki, które stało się lingua franca analizy danych (R Development Core Team, 2009). R ma bardzo potężny system graficzny, a także wieloplatformowe możliwości pakowania dowolnego kodu obliczeniowego. Setki dostępnych pakietów R, których liczba rośnie wykładniczo, wdrażają najnowocześniejsze metody obliczeniowe i odzwierciedlają stan badań w różnych dziedzinach. Pakiety R stały się odtwarzalnym narzędziem badawczym, ponieważ umożliwiają ponowne wykorzystanie i udostępnianie funkcji i algorytmów. Nie ma przeszkód do wdrożenia języka R na dużą skalę w publicznych chmurach i sieciach grid, ponieważ jest on objęty licencją GNU GPL (Chambers, 1998). Jednak R nie jest wielowątkowy i nie działa jako serwer. Jako język implementuje potężny system klasy S4, ale jako biblioteka R ma tylko niskopoziomowy, niezorientowany obiektowo interfejs programowania aplikacji (API).

Rozwój interfejsów (GUI) dla języka R pozostaje niestandaryzowany. Potencjał R jako obliczeniowego silnika zaplecza dla aplikacji i architektur zorientowanych na usługi nie został jeszcze w pełni wykorzystany. Chociaż baza użytkowników rośnie w szybkim tempie, to tempo wzrostu byłoby znacznie wyższe dzięki przyjaznemu i bogatemu środowisku pracy. Elastic-R wnosi do ekosystemu R

wszystkie te brakujące cechy, które mogą umożliwić zastosowanie go w wielu innych sytuacjach, na różne sposoby. Jego ambicja wykracza jednak daleko poza dostarczanie nowych narzędzi i ram. Rozszerzając logikę otwartości i rozszerzalności R, Elastic-R buduje środowisko, w którym wszystkie artefakty i zasoby obliczeniowe stają się „podłączone", a nie tylko komponent obliczeniowy (pakiet R). Rysunek 19.2 przedstawia kluczowe cechy Elastic-R. Środowisko pracy Java lub Ajax Elastic-R umożliwia naukowcowi, statystykowi, analitykowi finansowemu itp. łatwe połączenie (połączenie) możliwości synergii, opisanych w poniższych sekcjach:

Możliwość przetwarzania

Udostępniając prosty adres URL i dane uwierzytelniające, naukowiec łączy swoje środowisko pracy ze zdalnym silnikiem obliczeniowym opartym na języku R i uzyskuje dostęp do zasobów obliczeniowych, niezależnie od tego, czy jest to węzeł sieci Grid, maszyna wirtualna w chmurze, klaster czy własny laptop. Silnik jest niezależny od hostującego systemu operacyjnego i sprzętu. IaaS wymaga takiego mechanizmu, aby komputery stały się „jak elektryczność". EC2 pozwala użytkownikowi wybrać pojemność maszyny wirtualnej (liczba wirtualnych rdzeni, wielkość pamięci itp.), którą chciałby uruchomić. Środowisko pracy Elastic-R udostępnia ten wybór naukowcom w uproszczonej konsoli EC2. Stan silnika Elastic-R utrzymuje się do momentu zwolnienia zasobu obliczeniowego (zamknięcia maszyny wirtualnej, zabicia interaktywnego zadania Grid, zabicia procesu na serwerze fizycznym itp.). Naukowiec może odłączyć się od silnika i ponownie połączyć się z dowolnego miejsca: pobiera swoją sesję ze wszystkimi zmiennymi, funkcjami, grafiką, arkuszami kalkulacyjnymi itp.

Zdolność matematyczna i numeryczna

Uzyskując dostęp do sesji R i importując do swojego obszaru roboczego pakiety R związane z jego dziedziną problemową, naukowiec gromadzi funkcje i modele matematyczne potrzebne do przetwarzania danych i przekształcania ich w wiedzę i wgląd. Pakiet R może być wrapperem dowolnej biblioteki matematycznej napisanej w C, C++, FORTRAN itp. R może być uważany za uniwersalny framework dla kodu obliczeniowego i zestawów narzędzi obliczeniowych. Z poziomu swojej sesji R naukowiec może również zadzwonić do Scilab,5 Sage,7 Root,9 itd. i zwiększyć matematyczne możliwości swojego środowiska. Architektura rozszerzeń po stronie serwera pozwala każdemu na budowanie mostów Java, które łączą silnik Elastic-R z dowolnym oprogramowaniem. Takie mostki są dostępne dla Matlab10 i OpenOffice.11

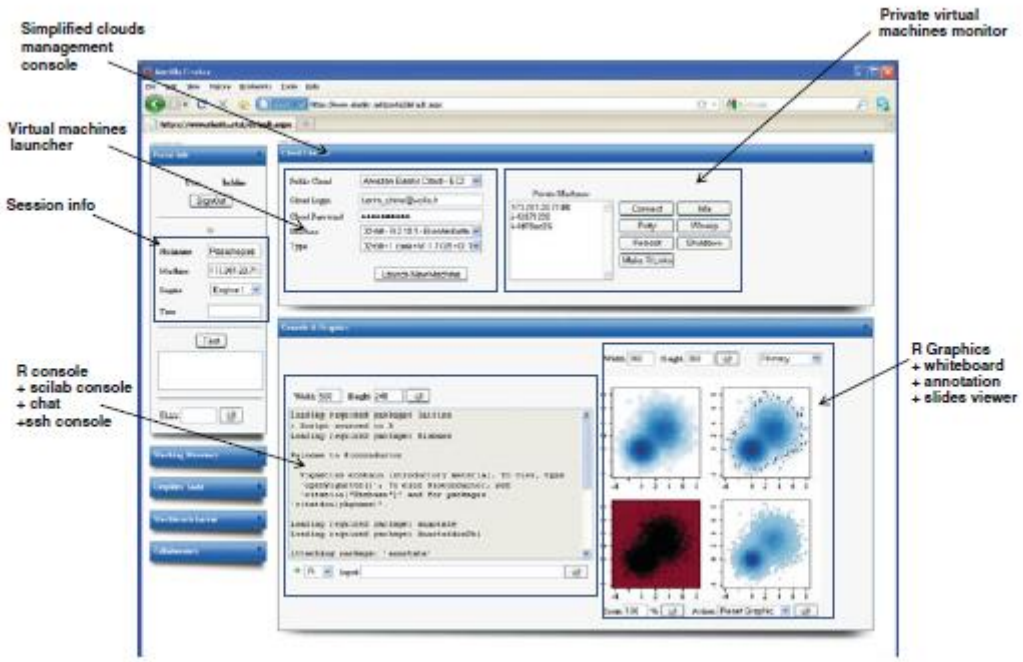
Zdolność orkiestracji

Język S zaimplementowany przez R jest jednym z najpotężniejszych języków, jakie kiedykolwiek stworzono do „programowania z danymi".6 Oprócz języka R, użytkownik może organizować zadania i kontrolować przepływ danych za pomocą Pythona i groovy. Interpretery dla tych języków skryptowych są wbudowane zarówno w silnik Elastic-R (po stronie serwera), jak i na stole warsztatowym (po stronie klienta). Pełne możliwości platformy są udostępniane za pośrednictwem front-endów SOAP i RESTful, a silnik Elastic-R może być pilotowany programowo z języka Java, Perl, C#, C++ itp. Dostarczane jest narzędzie aby umożliwić naukowcowi generowanie i wdrażanie usług SOAP Web Services, eksponując wybrane jego funkcje R (wygenerowane obliczeniowe usługi sieciowe). Mogą być używane jako węzły w środowiskach pracy. Węzły są dynamicznie połączone z sesją R naukowca, a przetwarzanie danych odbywa się w chmurze, jeśli w chmurze znajduje się silnik Elastic-R udostępniający usługę Web.

Możliwość interakcji

Widoki konsoli w środowiskach pracy Java i Ajax umożliwiają pełną kontrolę sesji R, a także użycie powłok Pythona, Groovy i Linux. Oprócz konsol, oba stanowiska pracy mają kilka wbudowanych

widoków dokowalnych, w tym zdalne przeglądarki katalogów do przeglądania, pobierania i wysyłania plików z i do katalogu roboczego zdalnego silnika, edytory kodu z podświetlaniem składni, przeglądarki pomocy, przeglądarki różnych plików formaty (PDF, SVG, HTML itp.), interaktywne urządzenia graficzne po stronie serwera z wbudowanymi funkcjami zmiany rozmiaru, powiększania, przewijania, śledzenia współrzędnych i adnotacji, inspektorzy danych, połączone wykresy, arkusze kalkulacyjne w pełni zintegrowane z funkcjami i danymi języka R. Architektura środowiska roboczego dla wtyczek pozwala każdemu tworzyć własne widoki i pulpity nawigacyjne, aby zwiększyć produktywność środowisk roboczych lub udostępniać modele statystyczne i numeryczne za pomocą prostych graficznych interfejsów użytkownika. Wszystkie widoki środowisk roboczych są oparte na współpracy: gdy więcej niż jeden użytkownik jest podłączony do tego samego silnika Elastic-R, działania jednego współpracownika są transmitowane do wszystkich pozostałych. Przykład środowiska pracy Elastic-R Java pokazano



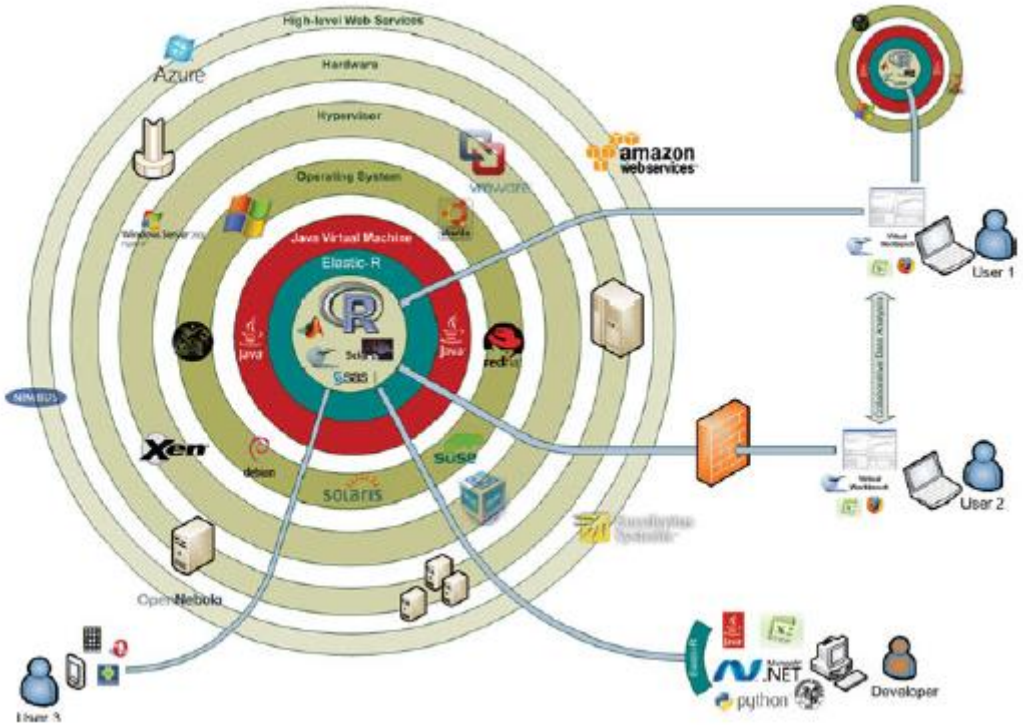
Zdolność wytrwałości

Katalog roboczy silnika obliczeniowego Elastic-R może znajdować się w lokalnym lub sieciowym systemie plików, a jego zawartość można łatwo synchronizować z serwerem FTP lub z Amazon S3. Gdy naukowiec używa uproszczonej konsoli EC2 do uruchomienia obrazu maszyny Amazon (AMI) z obsługą Elastic-R, magazyn Amazon Elastic Block Store (EBS) jest automatycznie dołączany do działającego AMI. Ten EBS staje się katalogiem roboczym wszystkich silników Elastic-R obsługiwanych przez AMI, a wszystkie pliki generowane przez naukowca, w tym serializacja obszarów roboczych, zawartość arkuszy kalkulacyjnych, wygenerowane usługi sieciowe itp., są przechowywane po zamknięciu AMI. EBS to także miejsce, w którym przechowywane są pakiety R zainstalowane przez naukowca. Pakiety te są udostępniane silnikom Elastic-R po uruchomieniu nowego AMI. Migawkę EBS może utworzyć naukowiec, który może zdecydować o udostępnieniu tej migawki innym użytkownikom EC2.

Elastic-R i Infrastructure-as-a-Service

Elastic-R może być używany na każdym typie infrastruktury. Jednak platforma nabiera pełnego wymiaru tylko wtedy, gdy jest używana w chmurze w stylu IaaS, niezależnie od tego, czy jest to Amazon EC2, czy prywatna chmura oparta na Eucalyptus,12 OpenNebula13 lub Nimbus.

Rysunek przedstawia rolę Elastic-R w środowisku infrastruktury jako usługi (koncentryczne okręgi).



Elastic-R otacza R (w samym środku) warstwami Java Object, do których można uzyskać zdalny dostęp z dowolnego miejsca. Utworzony silnik R (w kręgu wirtualnej maszyny Java) jest niezależny od systemu operacyjnego i do sprzętu. W takim przypadku działa na maszynie wirtualnej, która może być oparta na dowolnym systemie operacyjnym. Maszyna wirtualna korzysta z zasobów sprzętowych za pośrednictwem Hypervisora, a zarządzanie nią przez użytkownika końcowego (uruchamianie, wyłączanie itp.) odbywa się za pomocą warstwy zewnętrznej (API IaaS). „User1" i „User2" mogą łączyć się z silnikiem R i używać go wspólnie ze swoich środowisk roboczych Java, swoich środowisk roboczych Ajax lub arkuszy kalkulacyjnych Excel. Deweloper może korzystać z silnika R (jednego lub wielu, na jednej lub wielu maszynach wirtualnych), wywołując jego zdalne API (Java-RMI, SOAP, REST) ze swojej aplikacji internetowej ASP.NET, swojej aplikacji desktopowej Java, swoich skryptów Perl, Rys. 19.4 Elastic-R w środowisku IaaS, jego dodatek Excel itp. „Użytkownik 3" ze swojego urządzenia przenośnego (iPhone, telefon z systemem Android itp.) może użyć środowiska pracy Ajax, aby uzyskać dostęp do tego samego silnika R, co „ Użytkownik 1" i „Użytkownik 2" i wyświetlaj im jego dane, arkusze kalkulacyjne, slajdy itp. interaktywnie: środowisko pracy Ajax daje te same możliwości „Użytkownikowi 1", „Użytkownikowi 2" i „Użytkownikowi 3". Wszyscy mogą wydawać polecenia R, instalować i używać nowych pakietów, generować i wchodzić w interakcje z grafiką, przesyłać i przetwarzać pliki, pobierać wyniki itp.

Elementy składowe identyfikowalnej i odtwarzalnej obliczeniowej platformy badawczej

Elastic-R w chmurze w stylu IaaS zapewnia system, dzięki któremu można rejestrować środowisko obliczeniowe, dane i manipulacje danymi (skrypty, aplikacje). Mogą z nich korzystać recenzenci, współpracownicy i każdy, kto chce zbadać dane. Elastic-R zapewnia kompleksowe rozwiązanie do identyfikowalnych i powtarzalnych badań obliczeniowych. Migawki środowisk obliczeniowych można tworzyć jako obrazy maszyn wirtualnych (AMI). Migawki wersjonowanych bibliotek i katalogów

roboczych można tworzyć jako elastyczne magazyny bloków (EBS). Środowiska pracy Elastic-R Java i Ajax umożliwiają wszystkim naukowcom pracę z tymi migawkami (AMI + EBS) i łatwe ich tworzenie. Dzięki udostępnieniu identyfikatora AMI z obsługą Elastic-R, komplementarnych bibliotek obliczeniowych, identyfikatorów migawek EBS oraz identyfikatora migawki EBS katalogu roboczego (danych), które zostały użyte w jego badaniach, naukowiec umożliwia każdemu odbudowanie wszystkich danych i środowisko obliczeniowe wymagane do przetwarzania tych danych.

Podstawowe elementy platformy do nauczania statystyki i matematyki stosowanej

Oprócz tego, że jest darmowy i w większości open source, a zatem dostępny dla studentów i nauczycieli, Elastic-R zapewnia przyjazne dla edukacji funkcje, które do tej pory mogło oferować tylko oprogramowanie własnościowe (na przykład scentralizowane i kontrolowane wdrażanie naukowych środowisk obliczeniowych po stronie serwera) i umożliwia nowe scenariusze i praktyki w nauczaniu statystyki i matematyki stosowanej. Dzięki Elastic-R nauczyciele mogą ukryć złożoność R, Scilab, Matlab itp. za pomocą interfejsów użytkownika, takich jak wtyczki i arkusze kalkulacyjne Elastic-R. Są one bardzo łatwe do tworzenia i rozpowszechniania wśród uczniów. Interfejsy użytkownika zmniejszają złożoność środowiska uczenia się i trzymają początkujących uczniów z dala od stromych krzywych uczenia się R, Scilab lub Matlab. Po utworzeniu przez jednego nauczyciela interfejsy użytkownika mogą być udostępniane, ponownie wykorzystywane i ulepszane przez innych nauczycieli. Dedykowane repozytoria mogą być udostępniane w celu scentralizowania wysiłków i wkładu społeczności edukatorów oraz pomocy w dzieleniu się wiedzą zdobytą podczas korzystania z tego nowego środowiska. Można sobie wyobrazić wykorzystanie tych metod od szkół podstawowych po studia magisterskie. Nauczyciele mogą dostosowywać obrazy maszyn wirtualnych Elastic-R do konkretnych potrzeb swoich kursów i samouczków. Na przykład, po wybraniu najbardziej odpowiedniego obrazu, mogą dodać do niego brakujące pakiety R, wymagane pliki danych, zainstalować brakujące narzędzia itp. Nowy obraz można następnie udostępnić uczniom na kluczach USB lub udostępnić w IaaS chmura w stylu. W pierwszym przypadku, aby uruchomić środowisko pracy Elastic-R i połączyć się z silnikiem obliczeniowym na maszynie wirtualnej, uczestnicy kursu muszą mieć na swoich laptopach zainstalowaną Javę i odtwarzacz maszyny wirtualnej (na przykład darmowy odtwarzacz VMware). W drugim przypadku potrzebują tylko przeglądarki. Po raz kolejny maszyna wirtualna przygotowana przez jednego edukatora może być udostępniana, ponownie wykorzystywana i ulepszana przez innych edukatorów. Maszyna wirtualna jest w pełni samowystarczalna: kod potrzebny do uruchomienia środowiska roboczego lub przygotowane przez nauczyciela wtyczki mogą zostać dostarczone przez samo urządzenie wirtualne dzięki serwerowi kodu Elastic-R, który uruchamia się przy starcie. Interakcja między uczniem a SCE, a także wytwarzane artefakty są zapisywane w maszynie wirtualnej z włączoną funkcją Elastic-R. Nauczyciel może pobrać klucze USB używane przez uczniów (lub połączyć się z instancją maszyny wirtualnej w chmurze w stylu IaaS) i sprawdzić nie tylko ważność różnych uzyskanych wyników pośrednich, ale także ścieżkę, którą podążyli, aby uzyskać te wyniki. Możliwości współpracy w workbench otwierają również nowe perspektywy w nauczaniu rozproszonym. Nauczyciel może w każdej chwili połączyć się z SCE uczniów w dowolnym miejscu. Może przeglądać i aktualizować ich środowiska oraz zdalnie nimi kierować. Wspólne rozwiązywanie problemów staje się również możliwe i może być wykorzystywane jako wsparcie w nauce.

Elastic-R, narzędzie do e-nauki

Elastic-R to platforma e-Science, która zajmuje się niektórymi z najbardziej aktualnych przypadków użycia związanych z wykorzystaniem technologii informacyjno-komunikacyjnych (ICT) w badaniach i edukacji.

Obniżanie barier dostępu na żądanie

Infrastruktury komputerowe. Przejrzystość lokalna/zdalna

Ta sama aplikacja, Elastic-R workbench (rys. 19.6), ułatwia łączenie się z różnymi środowiskami lokalnie lub na zdalnych maszynach, niezależnie od tego, czy są to węzły sieci Grid, czy maszyny wirtualne w chmurze. Przełączanie się z jednego zasobu na inny (na przykład z jednej instancji maszyny wirtualnej w Amazon Elastic Compute Cloud na inną lub z interaktywnego zadania Grid w European Grid EGI do interaktywnego zadania w klastrze intranetowym) staje się tak proste, jak zastąpienie jednego adresu URL innym .

Radzenie sobie z potopem danych

Dane generowane przez nowoczesne narzędzia naukowe mogą stać się zbyt duże, aby można je było łatwo przenieść z jednej maszyny na drugą. Może to stanowić problem w przypadku dużych projektów współpracy. Analiza takich danych nie może być przeprowadzona tak, jak było do tej pory. Odpowiedzią na ten coraz bardziej dotkliwy problem jest przeniesienie obliczeń do danych i właśnie to umożliwia użytkownikom Elastic-R: ogólny silnik obliczeniowy może działać na dowolnej maszynie, która ma uprzywilejowaną łączność z maszyną do przechowywania danych lub na dużą skalę Baza danych. Tak jest w przypadku, gdy Elastic-R EC2 AMI jest używany do przetwarzania danych, które już znajdują się w Elastic Cloud Amazona. Użytkownik może podłączyć swój wirtualny warsztat (lub swoje skrypty za pomocą klientów Elastic-R SOAP) do silnika obliczeniowego, ustawić katalog roboczy na lokalizację danych (np. przez NFS) oraz przeglądać lub analizować dane za pomocą pakietów R/Scilab .

Włączanie współpracy w środowiskach obliczeniowych

Użytkownicy mogą łączyć się z tym samym zdalnym silnikiem i wspólnie pracować z danymi na dużą skalę, korzystając z nadawanych poleceń/grafik i wspólnych arkuszy kalkulacyjnych. Każdą komendę wydaną przez jednego z nich widzą wszyscy pozostali. Zsynchronizowane panele graficzne R umożliwiają im oglądanie tych samych grafik i wspólne opisywanie ich. Czat jest włączony. Połączone widoki wykresów oparte na zrefaktoryzowanym pakiecie iplots umożliwiają wspólne podświetlanie i malowanie kolorami na różnych grafikach o wysokiej interakcji.

Łatwe bramy naukowe

Interfejsy i portale internetowe umożliwiające naukowcom korzystanie ze sfederowanych, rozproszonych infrastruktur obliczeniowych do rozwiązywania problemów specyficznych dla danej domeny zawsze były trudne do opracowania, aktualizacji i utrzymania. Powinniśmy mieć proste frontendy. Elastic-R proponuje inny paradygmat tworzenia i dystrybucji takich front-endów do środowisk HPC/chmury z wtyczkami i arkuszami kalkulacyjnymi po stronie serwera

Wypełnianie luki między istniejącymi obliczeniami naukowymi

Środowiska i siatki/chmury

Po połączeniu środowiska roboczego użytkownika ze zdalnym silnikiem R/Scilab wbudowany serwer RESTful (lokalny przekaźnik http) umożliwia aplikacjom innych firm, takim jak emacs, OpenOffice Calc lub Excel, dostęp i korzystanie z silnika obsługującego Grid/chmurę. Na przykład dodatek do programu Excel umożliwia naukowcom korzystanie z pełnych możliwości programu Platformy Elastic-R i odtwarzaj funkcje arkuszy kalkulacyjnych Elastic-R z poziomu programu Excel. Dostępne jest również dwukierunkowe dublowanie modeli arkuszy kalkulacyjnych po stronie serwera w zakresach komórek Excela. Pozwala to użytkownikom przezwyciężyć niektóre błędy Excela (ograniczone możliwości analizy statystycznej, niedokładne obliczenia numeryczne na granicy podwójnej, niespójna identyfikacja

brakujących obserwacji...). Excel staje się front-endem wyboru dla zasobów gridowych/chmury, a następnie może stać się uniwersalnym warsztatem dla różnych nauk.

Wypełnianie luki między głównymi naukowymi środowiskami obliczeniowymi

Platforma posiada architekturę rozszerzeń po stronie serwera, która umożliwia tworzenie pomostów między zdalnym silnikiem obliczeniowym a dowolnym narzędziem innej firmy. Oprócz R i Scilab można zintegrować kilka powszechnie używanych środowisk (Matlab, Root, SAS itp.). Ponieważ R i Scilab działają w ramach tego samego procesu (ta sama wirtualna maszyna Javy), wymiana danych między nimi jest łatwa i bardzo szybka. Można to osiągnąć na przykład za pomocą interpretera Groovy dostępnego jako część zdalnego silnika. API SOAP można wywołać z dowolnego środowiska. Umożliwia użytkownikom SciPy na przykład pracę z silnikami Elastic-R w chmurze i wywoływanie funkcji R i Scilab.

Wypełnianie luki między głównymi naukowymi środowiskami obliczeniowymi a środowiskiem pracy

Elastic-R umożliwia automatyczną ekspozycję funkcji i pakietów R jako Web Services. Wygenerowane usługi sieci Web są łatwe do wdrożenia i mogą wykorzystywać silniki obliczeniowe zaplecza działające w dowolnej lokalizacji. Mogą być bezproblemowo zintegrowane jako węzły przepływów pracy i używane w środowiskach takich jak Knime, Taverna lub Pipeline Pilot. Mogą być bezstanowe (obliczenia wykonuje anonimowy pracownik R) lub stanowe (zarezerwowany pracownik R i powiązany z identyfikatorem sesji jest używane i mogą być używane ponownie, dopóki sesja nie zostanie zniszczona). Statefulness rozwiązuje problem narzutu spowodowany transferem wyników pośrednich między węzłami przepływu pracy.

Uniwersalny zestaw narzędzi obliczeniowych do zastosowań naukowych

Struktury i narzędzia Elastic-R umożliwiają używanie R jako zorientowanego obiektowo zestawu narzędzi Java lub jako serwera RMI. Wszystkie standardowe obiekty R zostały zmapowane do Javy ,a zdefiniowane przez użytkownika klasy R mogą być zmapowane do Javy na żądanie. Funkcje R mogą być wywoływane z Javy tak, jakby były funkcjami Javy. Parametry wejściowe są dostarczane jako obiekty Java, a wynik wywołania funkcji jest pobierany jako obiekt Java. Wywołania funkcji R z Javy lokalnie lub zdalnie radzą sobie z lokalnymi i rozproszonymi obiektami R. Pełne możliwości platformy są udostępniane za pośrednictwem interfejsów SOAP i RESTful. Dostępnych jest kilka narzędzi i struktur pomagających w tworzeniu analitycznych aplikacji desktopowych/sieciowych oraz skalowalnych potoków analizy danych w dowolnym języku programowania (Java, C#, C++, Perl itp.)

Skalowalność dla zaplecza obliczeniowego

Elastic-R zapewnia strukturę puli dla zasobów rozproszonych (RPF), umożliwiającą wdrażanie pul silników obliczeniowych na heterogenicznych węzłach/wystąpieniach maszyn wirtualnych. Silniki te są zarządzane i używane za pośrednictwem prostego interfejsu API pożyczania/zwracania dla wielowątkowych aplikacji internetowych i usług internetowych, do przetwarzania rozproszonego i równoległego, do dynamicznego generowania treści w locie (wyniki analityczne, tabele i grafika w różnych formatach dla klientów cienkiej sieci) oraz wirtualizacji silników obliczeniowych w kontekście współdzielonych zasobów obliczeniowych. Silniki stają się niezależne od systemu operacyjnego hostingu. Dostępnych jest kilka narzędzi do programowego lub interaktywnego monitorowania pul i zarządzania nimi (interfejs użytkownika administratora). Platforma poolingu umożliwia przejrzyste cloudbursting: instancje maszyn wirtualnych Amazon EC2 obsługujące jeden lub wiele silników obliczeniowych mogą być uruchamiane lub wyłączane w celu skalowania w górę lub w dół w zależności od obciążenia, na przykład przy wdrażaniu wysoce skalowalnych aplikacji internetowych.

Łatwe przetwarzanie rozproszone

Aby rozwiązać mocno obliczeniowe problemy, istnieje potrzeba równoległego używania wielu silników. Dostępnych jest kilka narzędzi, ale są one trudne do zainstalowania i wykraczają poza techniczne umiejętności większości naukowców. Elastic-R rozwiązuje ten problem. Z poziomu głównej sesji języka R i bez instalowania dodatkowych zestawów narzędzi/pakietów możliwe staje się tworzenie łączy logicznych do zdalnych silników R/Scilab poprzez tworzenie nowych procesów lub łączenie się z istniejącymi w sieciach/chmurach . Łącza logiczne to zmienne, które umożliwiają użytkownikowi R/Scilab interakcję z silnikami zdalnymi. rlink.console, rlink.get, rlink.put umożliwiają użytkownikowi przesyłanie poleceń R do pracownika R/Scilab, do którego odwołuje się rlink, pobieranie zmiennej z obszaru roboczego pracownika R/Scilab do głównego obszaru roboczego R i przekazywanie zmiennej z główny obszar roboczy R do obszaru roboczego pracownika. Wszystkie funkcje można wywoływać w trybie synchronicznym lub asynchronicznym. Kilka linków rlink odwołujących się do silników R/Scilab działających w dowolnych lokalizacjach można wykorzystać do utworzenia logicznego klastra, który umożliwia skoordynowane korzystanie z kilku silników R/Scilab. Na przykład funkcja o nazwie cluster.apply wykorzystuje równolegle procesy robocze należące do klastra logicznego w celu zastosowania funkcji do danych języka R na dużą skalę.

Elastic-R, platforma aplikacji dla chmury

Elastic-R można rozszerzać za pomocą komponentów Java zarówno po stronie klienta (wtyczki), jak i po stronie serwera (rozszerzenia). Dzięki tym komponentom każdy może tworzyć i wdrażać swoją aplikację w chmurze bez konkretnej wiedzy o infrastrukturze.

Przetwarzanie w chmurze i cyfrowa solidarność

Społeczność FOSS ustanowiła oprogramowanie Open Source jako wiarygodną alternatywę dla oprogramowania zastrzeżonego i umożliwiła użytkownikom z krajów rozwijających się bezpłatny dostęp do narzędzi najwyższej jakości. Przetwarzanie w chmurze umożliwia dziś każdemu korzystanie przez Internet z maszyn wirtualnych o dowolnej mocy obliczeniowej, działających w infrastrukturze sfederowanej. Połączenie sponsorowanego dostępu do chmur publicznych i oprogramowania dostarczanego jako usługa przez Internet otwiera perspektywy zmniejszenia przepaści cyfrowej na bezprecedensową skalę i umożliwia rewolucyjne nowe scenariusze dzielenia się wiedzą i cyfrowej solidarności. Elastic-R jest pierwszą platformą programową, która ujawnia ogromny potencjał tej kombinacji w badaniach i edukacji. Daje to, co najlepsze z istniejących naukowych środowisk obliczeniowych i narzędzi do analizy danych, w ręce wszystkich, udostępniając je jako usługę w chmurze Amazon (EC2). Badacze, nauczyciele i studenci z krajów rozwijających się mogą korzystać ze standardowych przeglądarek internetowych i wolnych łączy internetowych do pracy na przykład z R lub Scilab. Na przykład afrykańscy naukowcy mieliby dostęp na żądanie nie tylko do maszyn z dowolną liczbą procesorów i dowolnej wielkości pamięci w celu prowadzenia badań, ale także do potencjalnie nieskończonych współdzielonych zasobów cyfrowych. Zasobami tymi mogą być na przykład wstępnie przygotowane i gotowe do uruchomienia obrazy maszyn wirtualnych z narzędziami badawczymi lub danymi publicznymi lub aplikacjami analitycznymi dostarczanymi przez światowych naukowców. Portal Elastic-R obniża bariery dla każdego, kto może korzystać z chmury. Zapewnia również mechanizm bezpiecznych tokenów cyfrowych, które mogą być dostarczane przez międzynarodowe organizacje i organizacje charytatywne na przykład afrykańskim naukowcom. Tokeny pozwalają naukowcom uruchamiać maszyny wirtualne na określoną liczbę godzin i wykorzystywać je do swoich badań. Elastic-R to także Wirtualne Środowisko Badawcze, które pozwala dowolnej liczbie rozproszonych geograficznie użytkowników na jednoczesną i wspólną pracę z tą samą maszyną wirtualną, tym samym narzędziem i tymi samymi danymi. Ułatwia naukowcom z krajów rozwijających się aktywniejsze zaangażowanie się w duże międzynarodowe współpracę, utrzymywanie kontaktów naukowych w czasie rzeczywistym z ich rówieśnikami w USA i Europie oraz uzyskanie dostępu do

danych naukowych i środowisk obliczeniowych wymaganych do przetwarzania te dane. Elastic-R jest wreszcie systemem e-learningowym opartym na chmurze w czasie rzeczywistym, który umożliwia nauczycielom-wolontariuszom z krajów rozwiniętych interaktywne nauczanie statystyki i matematyki afrykańskim studentom bez konieczności przebywania w Afryce.

Wnioski i kierunki na przyszłość

Tu opisano Elastic-R jako nowe środowisko, które ma potencjał do demokratyzacji chmury i przyspieszenia odtwarzalności badań obliczeniowych. Jego obecna dostępność i łatwy dostęp w Amazon Elastic Compute Cloud maksymalizuje jego szanse na przyjęcie i adopcję. Uczelnie, przemysł i instytucje edukacyjne skorzystałyby na pojawieniu się nowego środowiska interoperacyjności, współdzielenia i ponownego wykorzystywania artefaktów obliczeniowych. Tworzenie i udostępnianie narzędzi i zasobów analitycznych może stać się dostępne dla każdego (otwarta nauka). Międzynarodowy portal4 do przetwarzania na żądanie jest budowany przy użyciu różnych struktur dostarczanych przez Elastic-R i może stać się pojedynczym punktem dostępu do zwirtualizowanych SCE na serwerach publicznych i na urządzeniach wirtualnych, które są gotowe do użycia w różnych chmurach. Nie ma wątpliwości co do potrzeby większej użyteczności w krajobrazie obliczeniowym. Java, Xen, VMware, EC2, R i Elastic-R udowadniają, że cel uniwersalnego środowiska obliczeniowego dla nauki i dla każdego jest zdecydowanie w zasięgu ręki.