Leveraging Azure From R

Azure Spark and MPI Clusters from R

Doug Service
Stephen Weller
Daniel Hanson

July 3, 2016

Microsoft Machine Learning
Revolution Analytics

Outline

1. Introduction
2. Azure
3. MPI Cluster
4. Portfolio Optimization Demo

Introduction

Goals

Leverage Azure compute clusters from R to solve compute or data parallel finance problems faster

1. Login to Azure accounts with $200 spending limit you can use during and after the presentation
2. Run and review R demos on pre-configured R Server Spark and MPI compute clusters

Azure

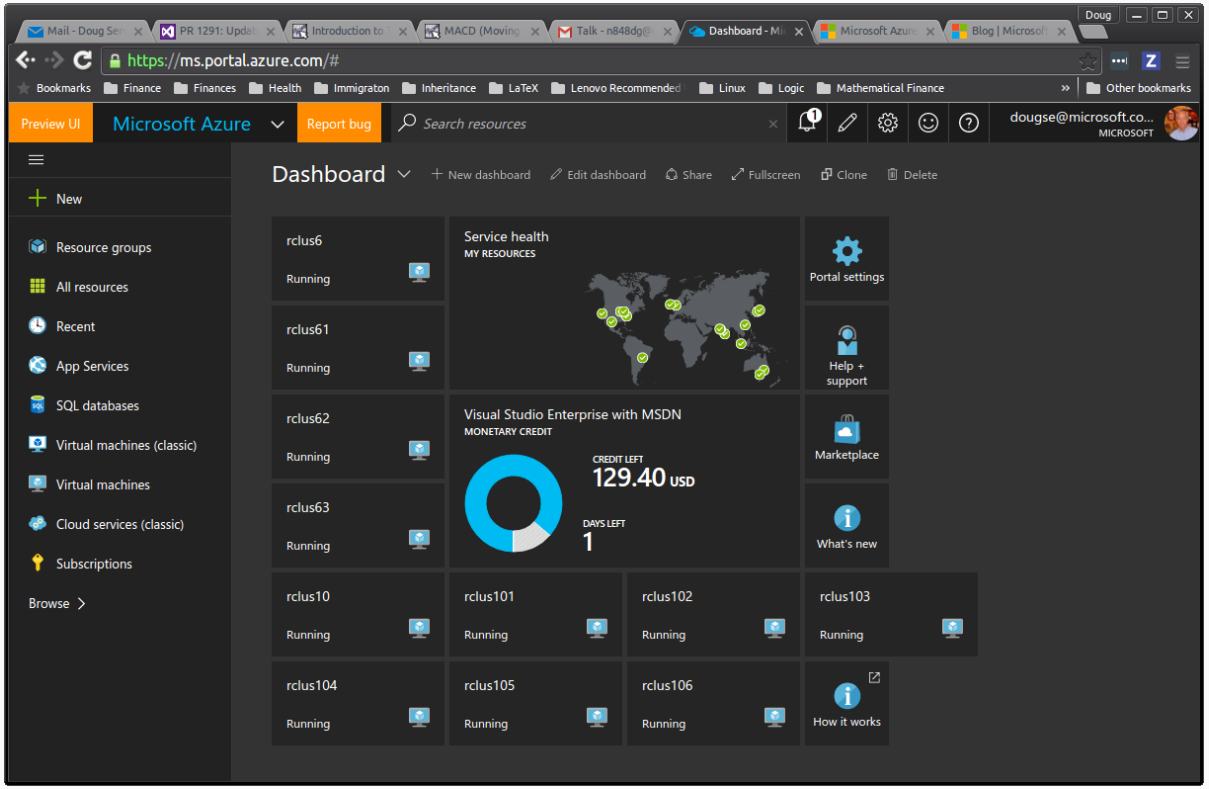
Advantages

Eliminates the expense of buying, maintaining, and continually upgrading a data center. Only pay for the resources you use.

Microsoft facility in Quincy Washington

Advantages

Build Spark, Hadoop, MPI compute clusters in Azure Portal, or languages such as Bash, PowerShell, node.js, or C# to access from R



Azure is a collection of integrated cloud services

- Compute - virtual machines (VMs)
  - Linux: Ubuntu, Redhat, CentOS...
  - Windows: Windows Server, Windows Enterprise...
- Networking - connect VMs
  - Internal virtual network
  - Public IP address and domain name
- Database - deploy to VMs
  - Oracle, OrientDB, Redis, SQL Server, MySQL
- Data Analytics - pre-configured
  - HDInsight, Stream Analytics, Cloudera
- Storage

MPI Cluster

MPI Cluster

Four virtual machines

All Nodes: desktop + worker

- Ubuntu Server 16.04
- Open message passing interface (OpenMPI)
- Open secure shell (OpenSSH)
- Network file system (NFS)
- R plus packages

Desktop node

- Ubuntu Mate Cloudtop desktop
- X remote desktop protocol (XRDP)
- Visual Studio Code editor
- Sublime Text 3 editor

MPI Cluster

R Packages

- foreach
- doMPI
- Rmpi

Gotchas

- rsh (ssh) must work reciprocally from all nodes, requires both public and private SSH key files on every node
- Development R scripts must be on all nodes in same location, best solution exports working directory on desktop node to compute nodes via Network File System (NFS)
- High performance configuration uses desktop in cloud due to high speed network connections to worker nodes

Portfolio Optimization Demo

S&P 500 Portfolio Optimization

Algorithm

- Select top 30% of stocks in each S&P index sector
  Industrials, Health Care, Information Technology etc.
- Form uniformly drawn random portfolios of 30 stocks
- Perform a minimum CVaR analysis on every portfolio
- Select the portfolio with the highest return
- Generate the efficient frontier for highest return portfolio

S&P 500 Portfolio Optimization

Optimization Run Time

Transport	Machines	Threads	Time (mins)	Script
None	1	1	4.3162	`RunPortST.sh`
MPI	1	4	1.6641	`RunPortMT.sh`
MPI	4	1	1.4296	`RunPortMPI.sh`

`RunAalysis.sh` - Generates analysis report

S&P 500 Portfolio Optimization

Demo Directory

`/nfs/mpidemos/rfinance/RAzureCluster/demo/portfolioOptimization`

Demo Files

- PortfolioMPI.R - portfolio optimization
- PortfolioMPIResults.R - generates optimization report

Using foreach

```
eres <- foreach(cdx=1:nnode,.packages='fPortfolio') %dopar% {
  # Get the combinations for the current node.
  ncmbs <- cmbs[,rngs[cdx,1]:rngs[cdx,2]]
  ret <- list()
  for (idx in 1:ncol(ncmbs)) {
    ret <- c(ret,list(list(Cmb=cmbs[,idx],
                Stats=calcMinCVaRPort(spxret.ts[,ncmbs[,idx]]))))
  }
  return(ret)
}
```

S&P 500 Portfolio Optimization

Review Portfolio optimization output