Data Analysis & Programming CA 2

Cesar Marrades Cortés – x00150376

Contents

[1. Scope 1](#_Toc500879591)

[2. Key Specs 1](#_Toc500879592)

[Optional Extra features 1](#_Toc500879593)

[3. Script Details 1](#_Toc500879594)

[Package Function Usages 2](#_Toc500879595)

[4. Usage 2](#_Toc500879596)

[Remarks: 2](#_Toc500879597)

[Arguments vs Default days 2](#_Toc500879598)

[Data Source 2](#_Toc500879599)

[Optimized for number of cores 2](#_Toc500879600)

[Validations 2](#_Toc500879601)

[Cluster Type and Redirection 3](#_Toc500879602)

[Tests 3](#_Toc500879603)

# Scope

The document below provides an overview of the script submitted for the assignment.

# Key Specs

* Identify stocks whose average daily gain in a certain time period is higher than the overall average daily gain of the entire stock exchange in that time
* Print out names of those stocks whose gain is higher than average

### Optional Extra features

* Parametrized script to limit days
* Parallelization
* Out of memory data

# Script Details

The steps followed to solve the problem are:

1. Compute the average daily gain for all stocks.
2. Setup a cluster with N nodes equal to N number of cores.
3. Process each stock in parallel by computing avg through a SqlLite query
4. Output values greater than avg.

# Package Function Usages

|  |  |  |
| --- | --- | --- |
| **Package** | **Function** | **Comments** |
| RSQLite | dbGetQuery |  |
| Parallel | MakeCluster |  |
| ClusterExport | To make variables available |
| ClusterCall | To Execute function in all clusters |
| parSapply | To process all elements in stockNames array |

# Usage

Arguments can be specified as per specification:

*commandArgs <- function(...) 9*

*source(‘program.r’)*

Please note script assumes sqlite file is on the same path.

# Remarks:

### Connections on clusters

Connection is not shared across clusters.

RSQLite is shared across clusters with the clusterCall function.

Days and

### Arguments vs Default days

Number of days can be specified using “*commandArgs*”. Otherwise it will be defaulted to 5

### Data Source

RSQLite used as source for the data.

### Optimized for number of cores

The script creates a cluster of same number of cores by using the “*detectCores()”* function.

### Validations

Note there is no validation for stocks negative gains scenario (*“Invest” maybe should be the option in this case*).

### Cluster Type and Redirection

Issues found with empty “outfile” parameter when assigning the type of cluster. Redirecting to text file works, but to meet specs

Cluster type is not specified on “makeCluster”, so prints could be redirected to standard output specifying “empty string” as “outfile”.



### Tests

“Manual” tests have been run to ensure single threaded and parallelized solution return same values.

