Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 56 (2015) 145 – 149

Procedia Computer Science

The 12th International Conference on Mobile Systems and Pervasive Computing (MobiSPC 2015)

Integrating of data using the Hadoop and R

Raissa Uskenbayeva[a], Abu Kuandykov[a], Young Im Cho[b], Tolganay Temirbolatova[a,*], Saule Amanzholova[c], Dinara Kozhamzharova[c,*]

[a]*IITU, 34A, Manas Str., Almaty, 050000,Kazakhstan*
[b]*Gachon University, 1342 Sungnam Daero, Sujung-gu, Seoul, 461-701, Korea*
[c]*KazNTU, 22 Satpayev Str., Almaty, 050013, Kazakhstan*

Abstract

The article offers a data integration model, which must be supported by a unified view of disparate data sources, management of integrity constraints, management of data manipulation and query executing operations, matching data from various sources, the ability to expand and set up new data sources. The proposed approach is the integration of Hadoop-based data and R, which is popular for processing statistical information. Hadoop database contains libraries, Distributed File System (HDFS), and resource management platform and implements a version of the MapReduce programming model for processing large-scale data. This model allows us to integrate various data sources at any level, by setting arbitrary links between circuit elements, constraints and operations.

Peer-review under responsibility of the Conference Program Chairs

Keywords: R, Big Data, Hadoop, Rhipe, RHadoop, Streaming.

1. Main text

The problem of integration of Big data from a variety and in particular, independent sources, combining heterogeneous data (without prejudice to their autonomy) seems to be quite relevant, especially in the context of the integration processes taking place at present in the world.

* Corresponding authors. Tel.: +7-701-522-52-69, +7-702-887-0055,
E-mail address: ttemirbolatova@gmail.com, dinara887@gmail.com.

Peer-review under responsibility of the Conference Program Chairs
doi:10.1016/j.procs.2015.07.187

It should be noted that today only the basic outlines of a generalized approach to the integration of data are outlined, especially when it comes to integration at the data level.

In this case, it is good to use Apache Hadoop and programming language R, which can ensure the integrity of data during the integration.

Apache Hadoop is base which is implemented by open source software written in Java for distributed storage and distributed processing of very large data sets on computing clusters, built from commodity hardware.

All modules in Hadoop are designed with respect of fundamental assumption that hardware failures (individual machines, or machines racks) are commonplace and, therefore, should be automatically processed within the software.

R – is a programming language for statistical data processing and graphics, as well as free software environment for computing open source in the GNU project.

The language was created as the same as language S, developed at Bell Labs, and is an alternative implementation, although between languages have significant differences, but for the most part in the language of the code runs on S in the R environment.

We will present three approaches to integrate R and Hadoop: R and Streaming, Rhipe and RHadoop. There are also other approaches to integrate R and Hadoop. For example RODBC/RJDBC could be used to access data from R.

The general structure of the analytics tools integrated with Hadoop can be viewed as a layered architecture presented in Fig. 1.

The first layer is the hardware layer – it consists in a cluster of (commodity) computers. The second layer is the middleware layer – Hadoop. It manages the distributions of the files by using HDFS and the MapReduce jobs. Then it comes a layer that provides an interface for data analysis. At this level we can have a tool like Pig which is a high-level platform for creating MapReduce programs using a language called Pig-Latin. We can also have Hive which is a data warehouse infrastructure developed by Apache and built on top of Hadoop. Hive provides facilities for running queries and data analysis using an SQL-like language called HiveQL and it also provides support for implementing MapReduce tasks.

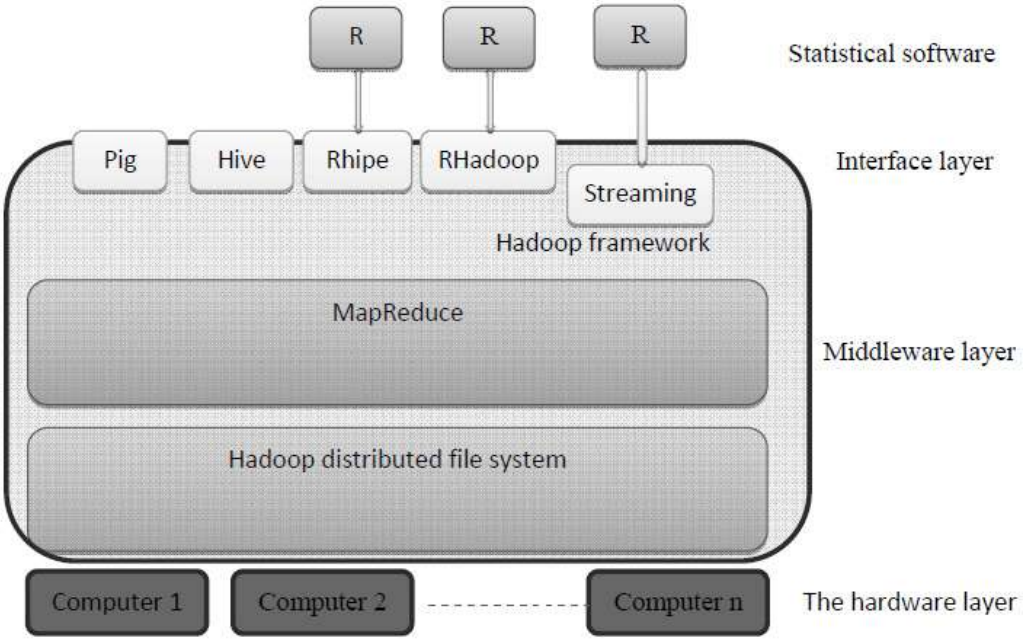

Fig. 1. Hadoop and data analysis tools

Besides these two tools we can implement at this level an interface with other statistical software like R. We can use Rhipe or Rhadoop libraries that build an interface between Hadoop and R, allowing users to access data from the Hadoop file system and write their own scripts for implementing Map and Reduce jobs, or we can use Streaming that is a technology integrated in Hadoop.

2. R and Streaming

Streaming is a technology integrated in the Hadoop distribution that allows users to run Map/Reduce jobs with any script or executable that reads data from standard input and writes the results to standard output as the mapper or reducer. This means that we can use Streaming together with R scripts in the map and/or reduce phase since R can read/write data from/to standard input. In this approach there is no client-side integration with R because the user will use the Hadoop command line to launch the Streaming jobs with the arguments specifying the mapper and reducer R scripts.

A command line with map and reduce tasks implemented as R scripts would look like the following (see Fig. 2.):

```
$ ${HADOOP_HOME}/bin/Hadoop jar ${HADOOP_HOME}/contrib/streaming/*.jar \ -inputformat
org.apache.hadoop.mapred.TextInputFormat \ -input input_data.txt \
-output output \
-mapper /home/tst/src/map.R \
-reducer /home/tst/src/reduce.R \
-file /home/tst/src/map.R \
-file /home/tst/src/reduce.R
```

Fig. 2. An example of a Map-reduce task with R and Hadoop integrated by Streaming framework

The integration of R and Hadoop using Streaming is an easy task because the user only needs to run Hadoop command line to launch the Streaming job specifying the mapper and reducer scripts as command line arguments. This approach requires that R should be installed on every DataNode of the Hadoop cluster but this is simple task.

The licensing scheme need for this approach implies an Apache 2.0 license for Hadoop and a combination of GPL-2 and GPL-3 for R.

3. RHIPE

Rhipe stands for “R and Hadoop Integrated Programming Environment” and is an open source project that provides a tight integration between R and Hadoop. It allows the user to carry out data analysis of big data directly in R, providing R users the same facilities of Hadoop as Java developers have. The software package is freely available for download at www.datadr.org [5].

The installation of the Rhipe is somehow a difficult task. On each DataNode the user should install R, Protocol Buffers and Rhipe and this is not an easy task: it requires that R should be built as a shared library on each node, the Google Protocol Buffers to be built and installed on each node and to install the Rhipe itself. The Protocol Buffers are needed for data serialization, increasing the efficiency and providing interoperability with other languages.

The Rhipe is an R library which allows running a MapReduce job within R. The user should write specific native R `map` and `reduce` functions and Rhipe will manage the rest: it will transfer them and invoke them from map and reduce tasks. The map and reduce inputs are transferred using a Protocol Buffer encoding scheme to a Rhipe C library which uses R to call the map and reduce functions [6]. The advantages of using Rhipe and not the parallel R packages consist in its integration with Hadoop that provides a data distribution scheme using Hadoop distributed file system across a cluster of computers that tries to optimize the processor usage and provides fault tolerance.

The general structure of an R script that uses Rhipe is shown in Fig. 3. and one can easily note that writing such a script is very simple.

```
library (Rhipe)
rhinit (TRUE, TRUE);
map<-expression ({lapply (map.values, function (mapper)...)})
reduce <-expression (
pre = {...},
reduce = {...},
post = {...},
x <- rhmr (
map=map, reduce=reduce,
          ifolder=inputPath,
          ofolder=outputPath,
          inout=c ('text', 'text'),
          jobname='a job name'))
rhex (z)
```

Fig. 3. The structure of an R script using Rhipe

Rhipe let the user to focus on data processing algorithms and the difficulties of distributing data and computations across a cluster of computers are handled by the Rhipe and library and Hadoop.

4. RHADOOP

RHadoop is an open source project developed by Revolution Analytics that provides client-side integration of R and Hadoop [3].

Setting up RHadoop is not a complicated task although RHadoop has dependencies on other R packages. Working with RHadoop implies to install R and RHadoop packages with dependencies on each Data node of the Hadoop cluster. RHadoop has a wrapper R script called from Streaming that calls user defined map and reduce R functions. RHadoop works similarly to Rhipe allowing user to define the `map` and `reduce` operation. A script that uses RHadoop looks like the following (see Fig. 4.):

```
    library (rmr)
    map<-function (k, v) { ...}
    reduce<-function (k, vv) {...}
    mapreduce (
input ="data.txt",
        output="output",
        textinputformat =rawtextinputformat,
map = map, reduce=reduce)
```

Fig. 4. The structure of an R script using RHadoop

It should be noted that `rmr` makes the client-side R environment available for `map` and `reduce` functions. The licensing scheme needed for this approach implies an Apache 2.0 license for Hadoop and RHadoop and a combination of GPL-2 and GPL-3 for R.

5. Conclusions

Official statistics is increasingly considering big data for building new statistics because its potential to produce more relevant and timely statistics than traditional data sources. One of the software tools successfully used for storage and processing of big data sets on clusters of commodity hardware is Hadoop. In this paper we presented three ways of integrating R and Hadoop for processing large scale data sets: R and Streaming, Rhipe and RHadoop. We have to mention that there are also other ways of integrating them like ROBDC, RJBDC or Rhive but they have

some limitations. Each of the approaches presented here has benefits and limitations. While using R with Streaming raises no problems regarding installation, Rhipe and RHadoop requires some effort in order to set up the cluster. The integration with R from the client side part is high for Rhipe and Rhadoop and is missing for R and Streaming. Rhipe and RHadoop allows users to define and call their own `map` and `reduce` functions within R while Streaming uses a command line approach where the `map` and `reduce` functions are passed as arguments. Regarding the licensing scheme, all three approaches require GPL-2 and GPL-3 for R and Apache 2.0 for Hadoop, Streaming, Rhipe and RHadoop.

We have to mention that there are other alternatives for large scale data analysis: Apache Mahout, Apache Hive, commercial versions of R provided by Revolution Analytics, Segue framework or ORCH, an Oracle connector for R but Hadoop with R seems to be the most used approach [8]. For simple Map-Reduce jobs the straightforward solution is Streaming but this solution is limited to text only input data files. For more complex jobs the solution should be Rhipe or RHadoop.

References

1. R.Uskenbayeva, Y.Chinibayev, A.Kassymova, T.Temirbolatova, K. Mukhanov. Technology of integration of diverse databases on the example of medical records// 2014 14th International Conference on Control, Automation and Systems (ICCAS 2014) Oct. 22-25, 2014 in KINTEX, Gyeonggi-do, Korea.
2. R.Uskenbayeva. S.T.Amanzholova, G.I. Khasenova, T.Temirbolatova. Analysis and Localization of Performance Degradation Incidents of Distributed Computer Systems// Smart-government: Proceeding of the International Scientific-Practical Conference. - 2014. - P. 75-81.
3. Bektemyssova G., Chinibayev Y., Temirbolatova T., Uskenbaeva R. Recovery and visualization of 3D models as a part of integrated database The 12th INTERNATIONAL CONFERENCE INFORMATION TECHNOLOGIES AND MANAGEMENT 2014 April 16-17, 2014, Information Systems Management Institute, Riga, Latvia
4. R.Uskenbayeva, G.Bektemyssova, T.Temirbolatova, A.Kassymova. Recursive decomposition as a method for integrating heterogeneous data sources 2015 15th International Conference on Control, Automation and Systems (ICCAS 2015) Oct. 13-16, 2015 in BEXCO, Busan, Korea (in press)
5. http://www.datadr.org // Divide and Recombine (D&R) with RHIPE Deep Analysis of Complex Big Data using the R Environment
6. http://www.revolutionanalytics.com // The revolution Analytics perspective on Big Data
7. White, T., (2012), Hadoop: The Definitive Guide, 3rd Edition, O'Reilly Media.
8. Yahoo! Developer Network, (2014), Hadoop at Yahoo!, available at http:// developer.yahoo.com/hadoop/, last accessed on 25th March, 2014.