llama Big Data Integration and Analysis

Authors:

Gaurav Kumar, Héctor Ugarte, Miguel Mármol, Tina Boroukhian

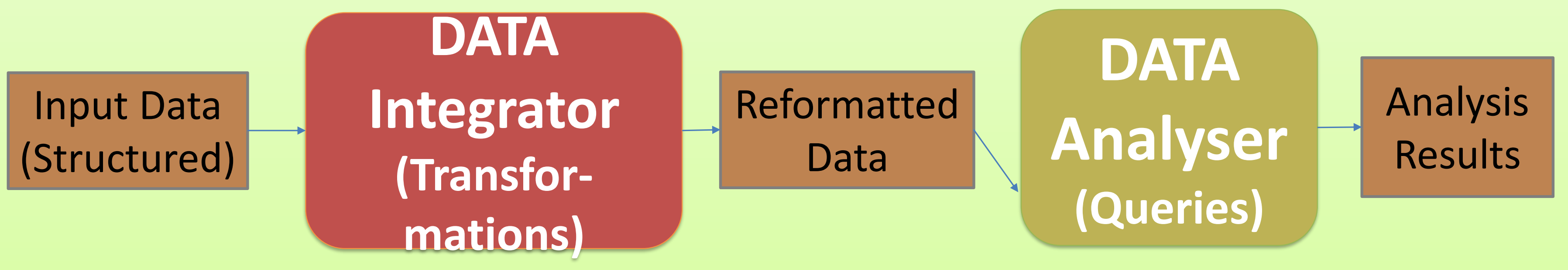
Summer Semester 2015

Project Overview

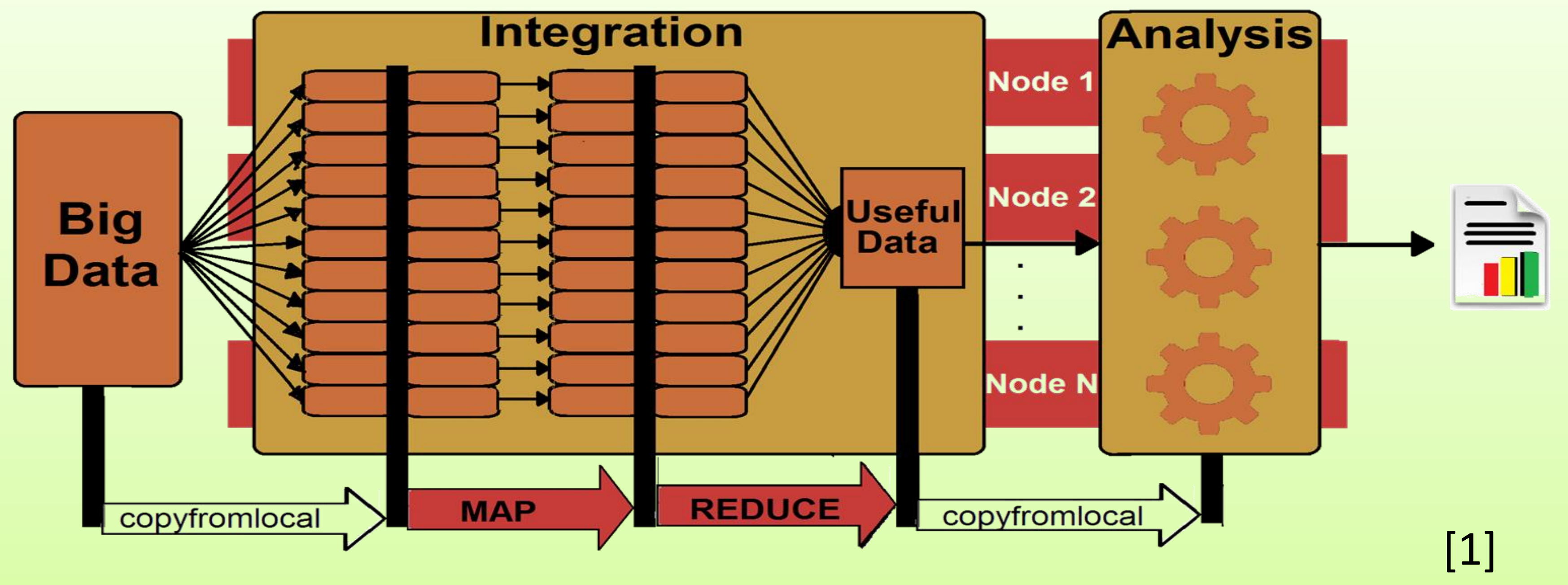
Llama is Data Integration and Analysis Platform made of two components:

1. Data Integrator: llama first loads structured plain files. Then data loaded is cleaned, reformatted and filtered using a series of user-selected transformations. Integration jobs can be specified in two ways: either via (1) a graphical User Interface, or (2) a script written in **IJSL**, for **I**ntegration **J**ob **S**pecification **L**anguage. If the first method is used, at the end, the job can be exported as an IJSL script that can be used in a later run.

2. Data Analyzer: Once ready, the new data is analyzed by means of SQL queries. The results can be stored for possibly further analysis.


Input Data (Structured) → DATA Integrator (Transfor-mations) → Reformatted Data → DATA Analyser (Queries) → Analysis Results

System Architecture



[1]

Implementation

Apache Hadoop is an open-source framework for distributed storage and processing of very large data sets on a of commodity hardware. [2]

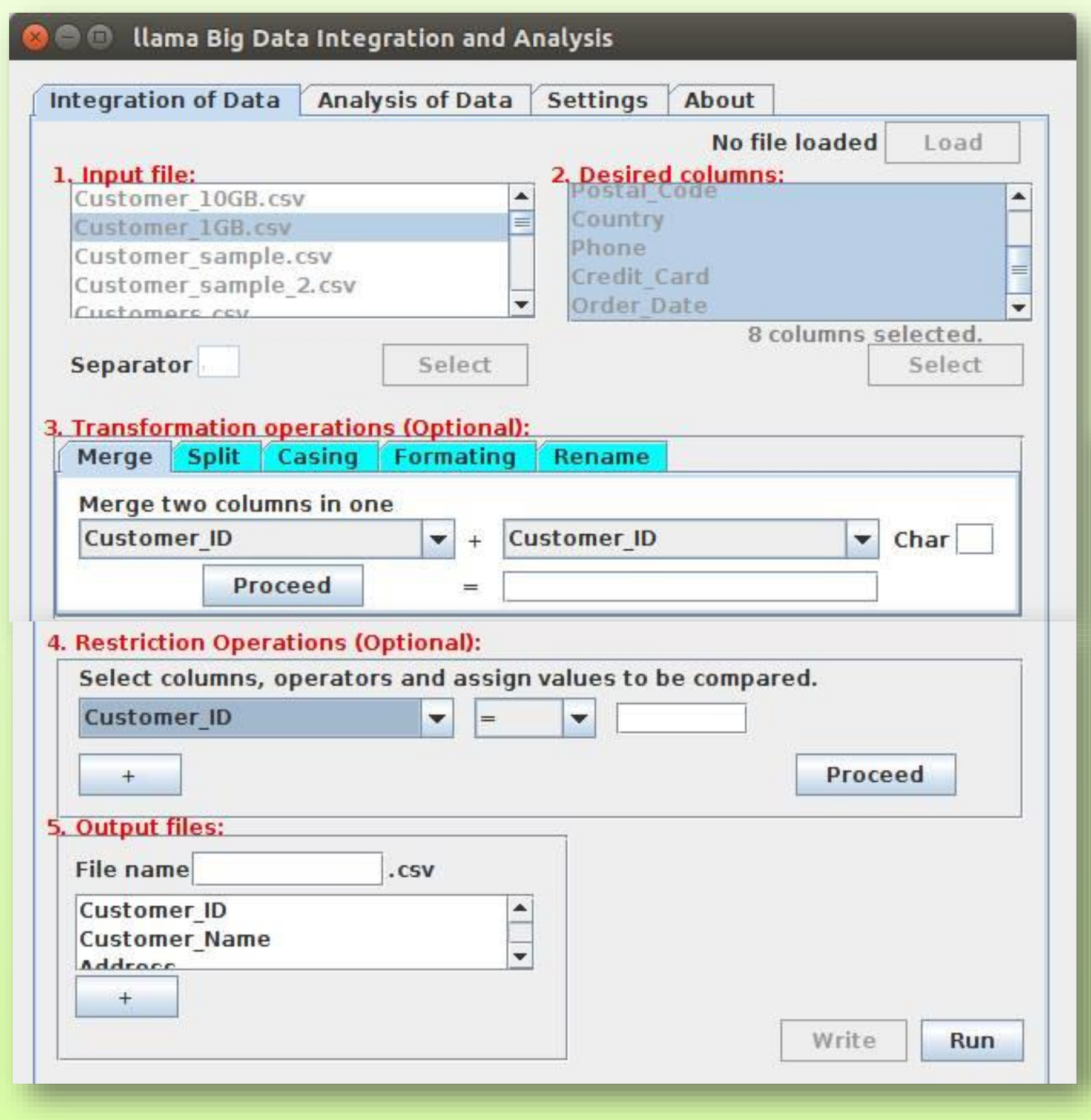
Tutorial for installation: http://pingax.com/install-apache-hadoop-ubuntu-cluster-setup/

Apache Spark is an open-source cluster computing framework. It employs the concept of RDD which are distributed units of data that resides primarily in memory, hence its high speed. [3]

Tutorial for installation: http://www.trongkhoanguyen.com/2014/11/how-to-install-apache-spark-121-in.html

Data Integrator

Input File Selection → Columns Selection → Transfor-mations Selection → Restriction Selection → Output File



No file loaded

1. Input file:
- Customer_10GB.csv
- Customer_1GB.csv
- Customer_sample.csv
- Customer_sample_2.csv
- Customers.csv

2. Desired columns:
- Postal_Code
- Country
- Phone
- Credit_Card
- Order_Date

8 columns selected.

Separator

3. Transformation operations (Optional):

Merge two columns in one

Customer_ID + Customer_ID Char

=

4. Restriction Operations (Optional):

Select columns, operators and assign values to be compared.

Customer_ID =

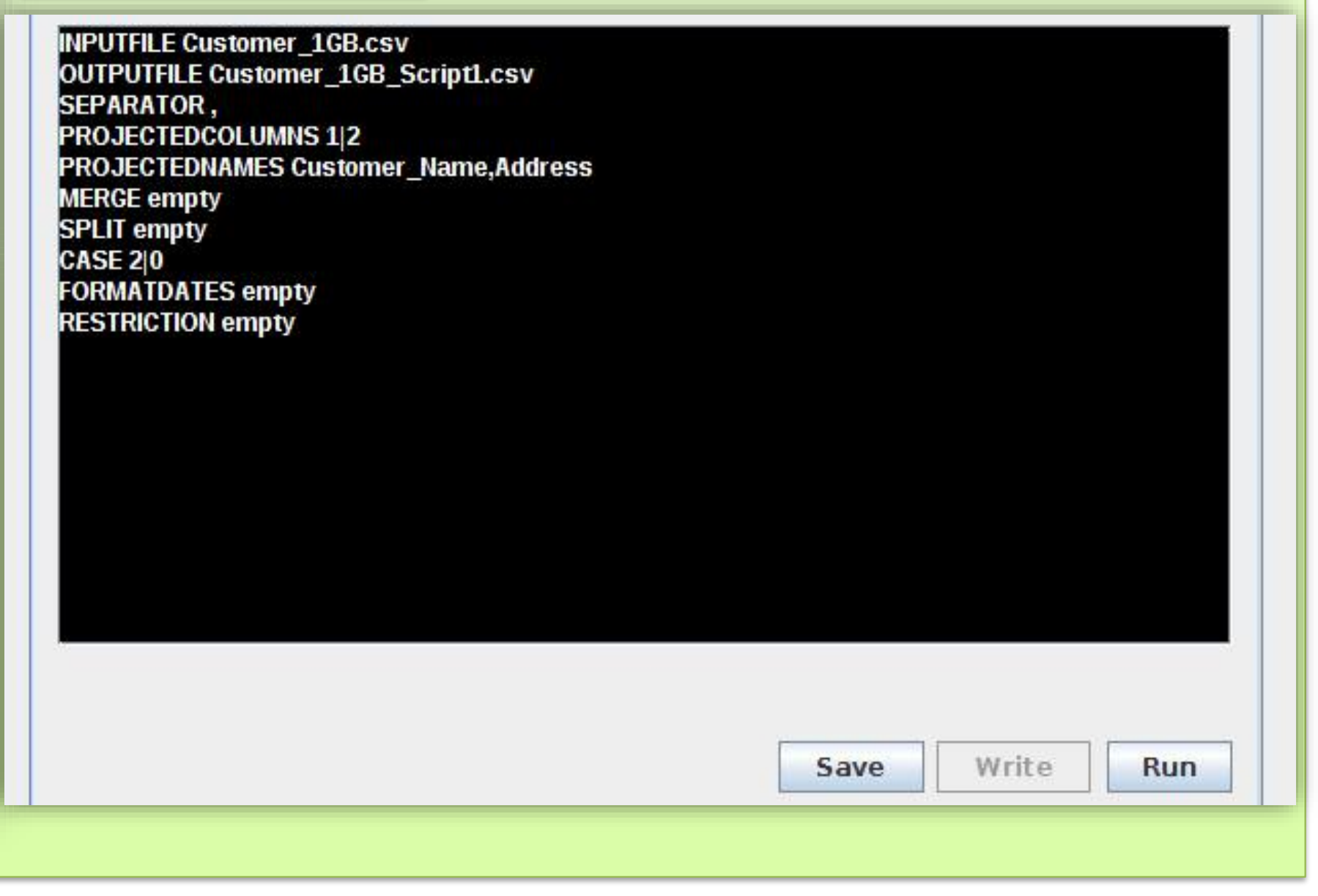
5. Output files:

File name .csv

- Customer_ID
- Customer_Name
- Address

(*) alternative flow

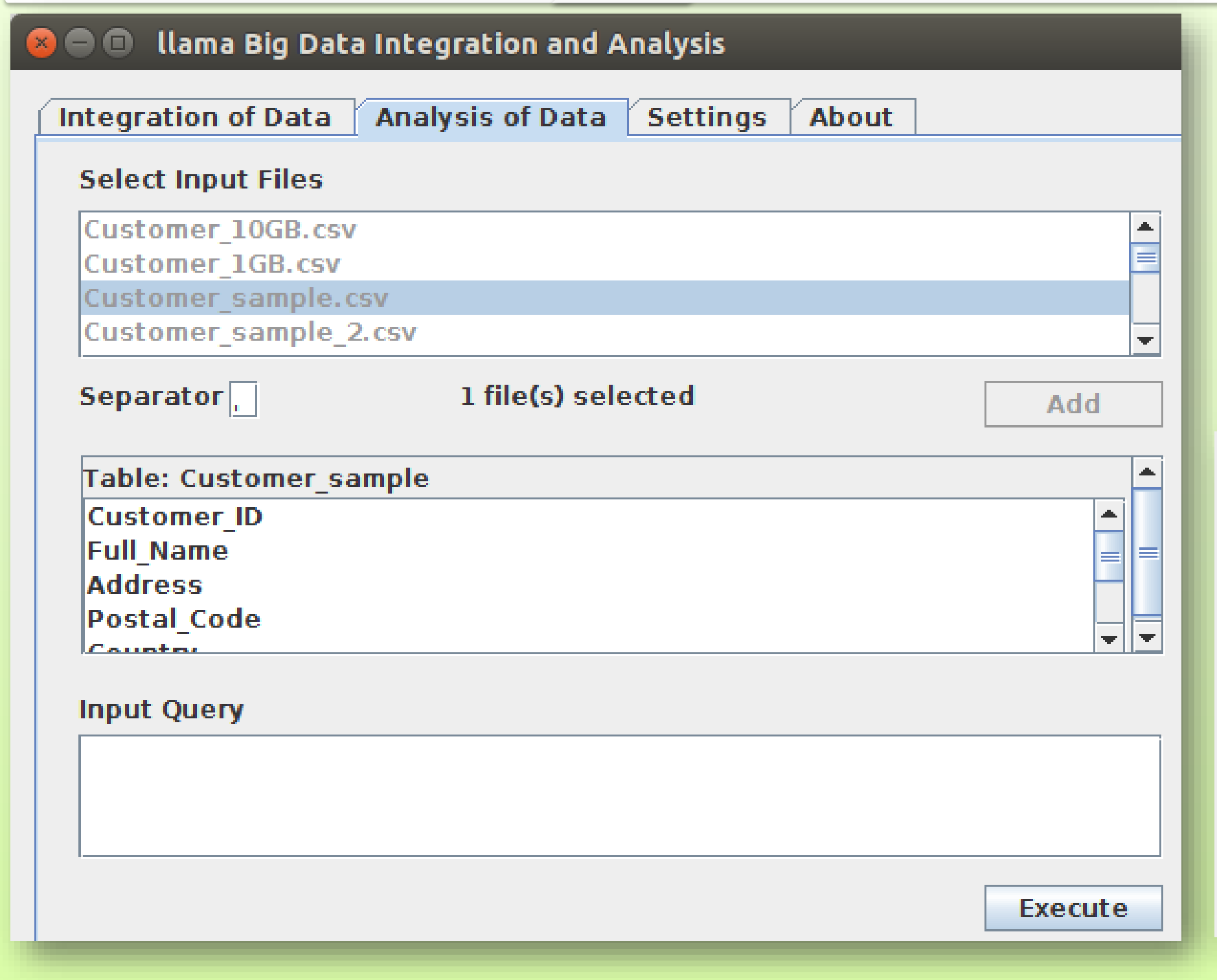
Write or Load IJSL Script



```
INPUTFILE Customer_1GB.csv
OUTPUTFILE Customer_1GB_Script1.csv
SEPARATOR ,
PROJECTEDCOLUMNS 1|2
PROJECTEDNAMES Customer_Name,Address
MERGE empty
SPLIT empty
CASE 2|0
FORMATDATES empty
RESTRICTION empty
```

Data Analyser

Input File Selection → Writing Query → Visualization & Saving Results



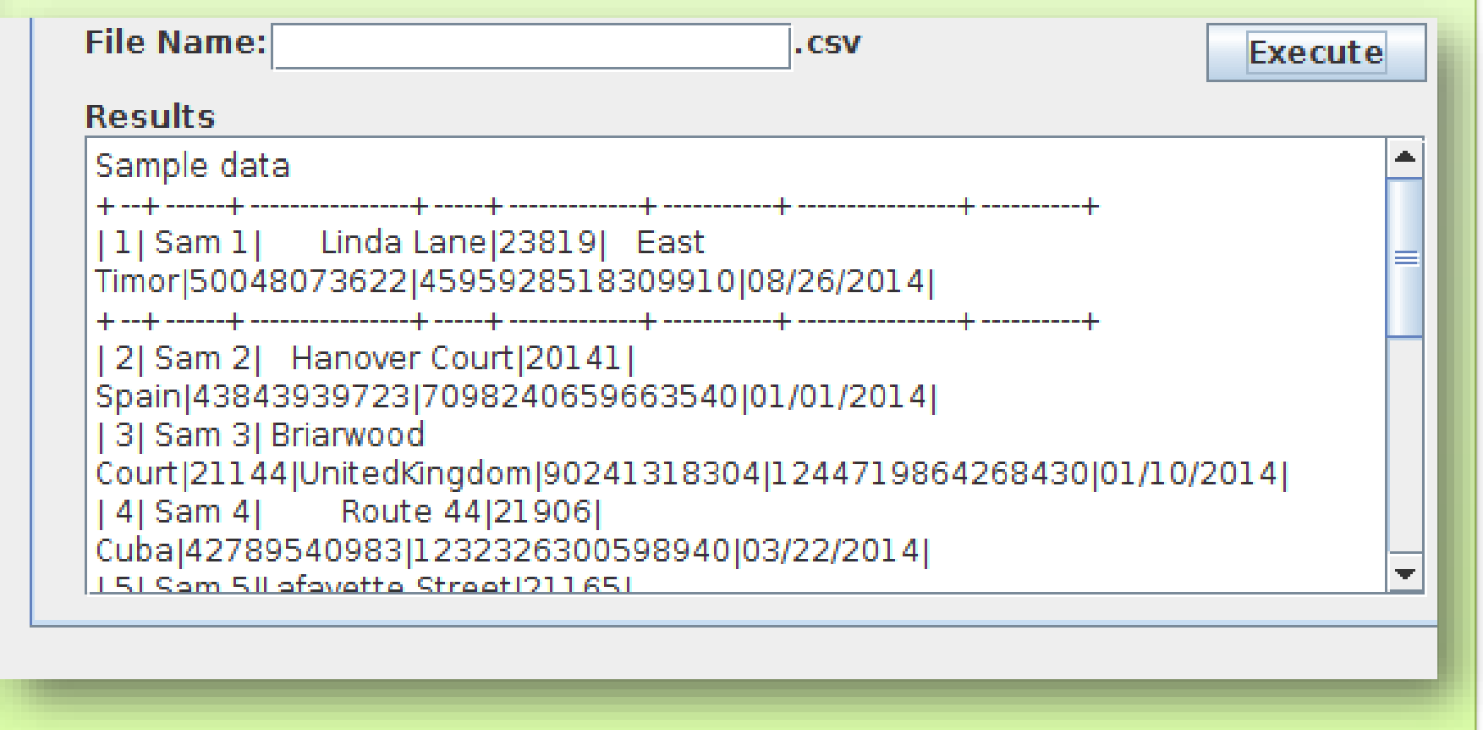
Select Input Files
- Customer_10GB.csv
- Customer_1GB.csv
- Customer_sample.csv
- Customer_sample_2.csv

Separator ,

1 file(s) selected

Table: Customer_sample
- Customer_ID
- Full_Name
- Address
- Postal_Code
- Country

Input Query

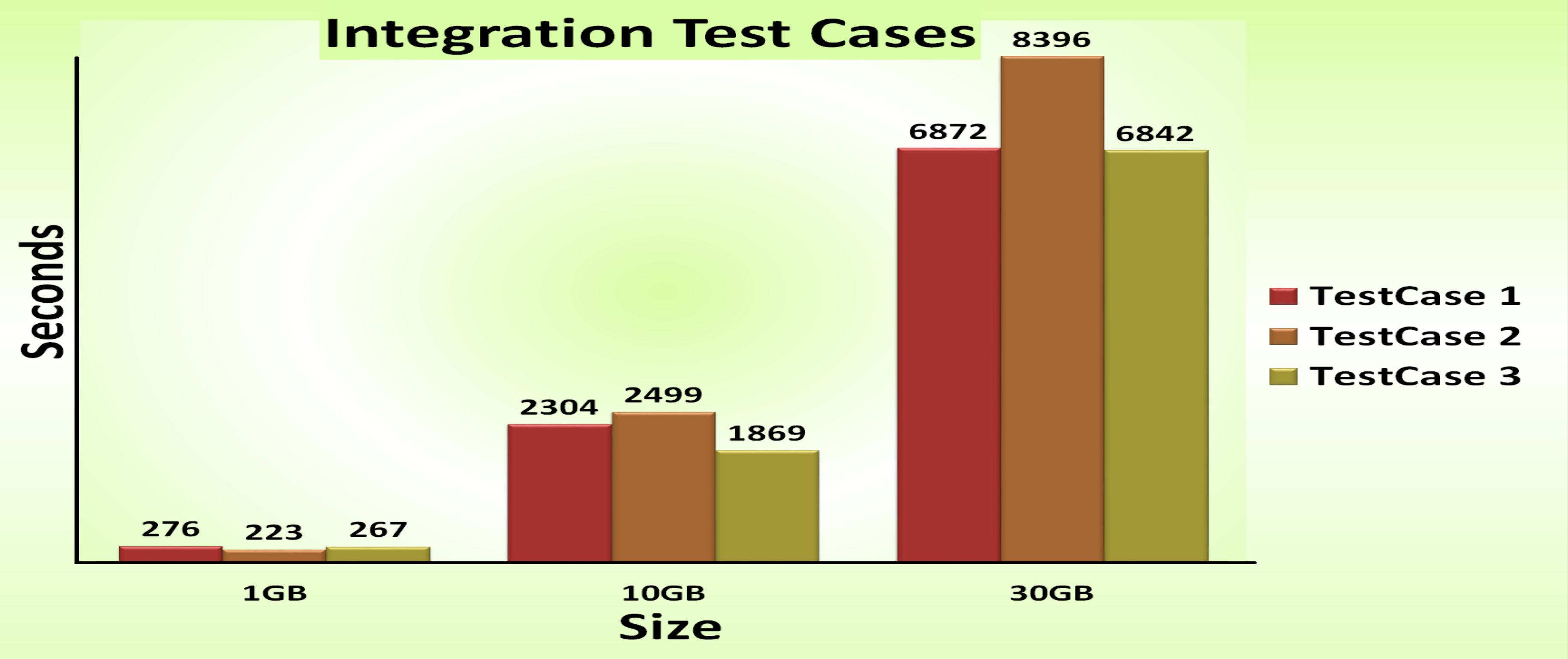


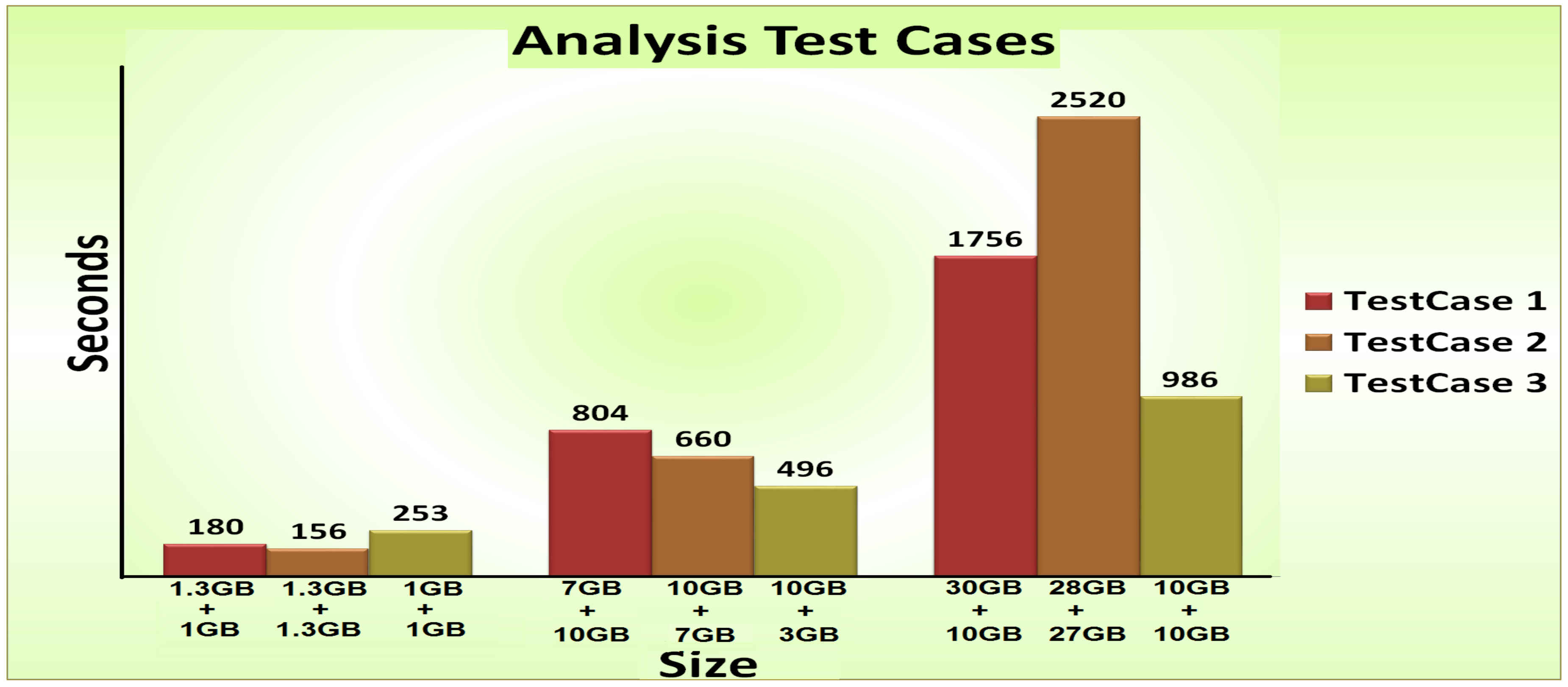
File Name: .csv

Results

```
Sample data
+--+------+---------------+-----+------------+----------+---------------+----------+
| 1| Sam 1|      Linda Lane|23819|   East
Timor|50048073622|4595928518309910|08/26/2014|
+--+------+---------------+-----+------------+----------+---------------+----------+
| 2| Sam 2|   Hanover Court|20141|
Spain|43843939723|7098240659663540|01/01/2014|
| 3| Sam 3| Briarwood
Court|21144|UnitedKingdom|90241318304|1244719864268430|01/10/2014|
| 4| Sam 4|        Route 44|21906|
Cuba|42789540983|1232326300598940|03/22/2014|
| 5| Sam 5|Lafayette Street|21165|
```

Evaluations





Live Demo

Two virtual machines hosted on EIS02

Local Access Master:
User: eis-groupb-master
Password: EIS2015

Hadoop User:
User: hduser
Password: hduser

Remote Access:
Partner ID: 572120720
Password: 268vxh

Local Access Node:
User:eis-groupb-slave
Password: EIS2015

Software installed:
- Ubuntu 14.04 LTS 64bit
- JDK 1.8
- Apache Hadoop 2.6
- Apache Spark 1.4

Remote Access Node:
Partner ID: 572193139
Password: 4d2g1h

Github Link:

[1] User Manual: https://docs.google.com/document/d/1Rhdg8mlKdiJpZoFBIyfrao1-czLz4FkjaBdnLxr8V60/edit

[2] Technical Report: https://docs.google.com/document/d/1kNYN4YxueQ0UbKZSeWWSwx5JEsuJFOX08XeEkN7a8S0/edit

References :

[1] http://www.glennklockwood.com/data-intensive/hadoop/mapreduce-workflow.png
[2] https://en.wikipedia.org/wiki/Apache_Hadoop
[3] https://en.wikipedia.org/wiki/Apache_Spark