llama Big Data Integration and Analysis

Authors:

Gaurav Kumar, Héctor Ugarte, Miguel Mármol, Tina Boroukhian

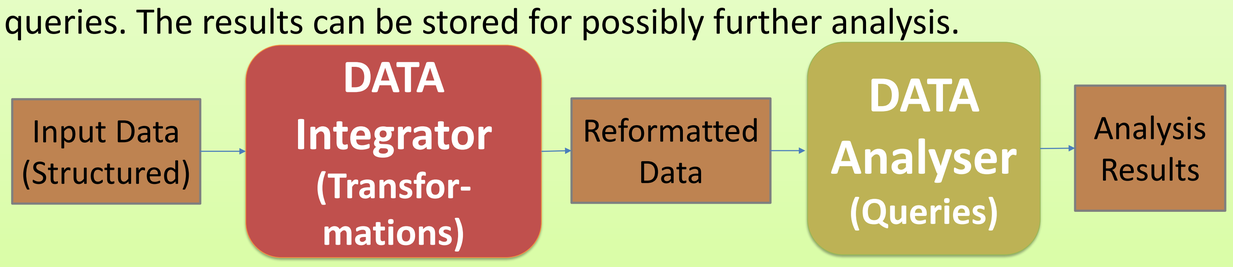
Summer Semester 2015

Project Overview

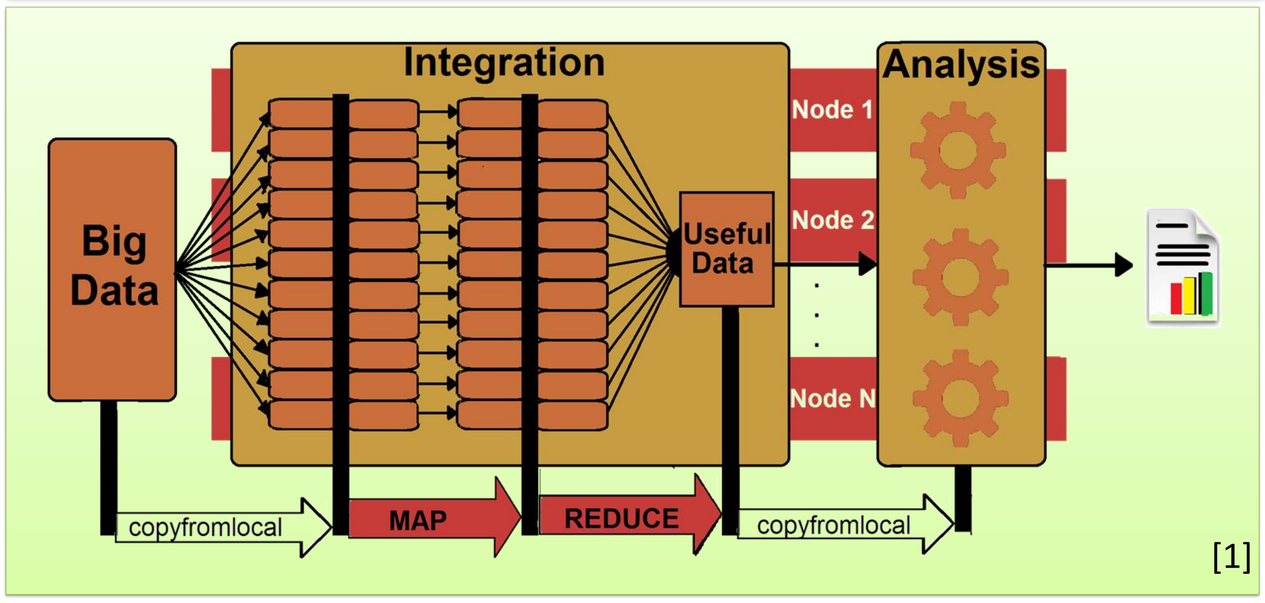
Llama is Data Integration and Analysis Platform made of two components:

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.



System Architecture



[1]

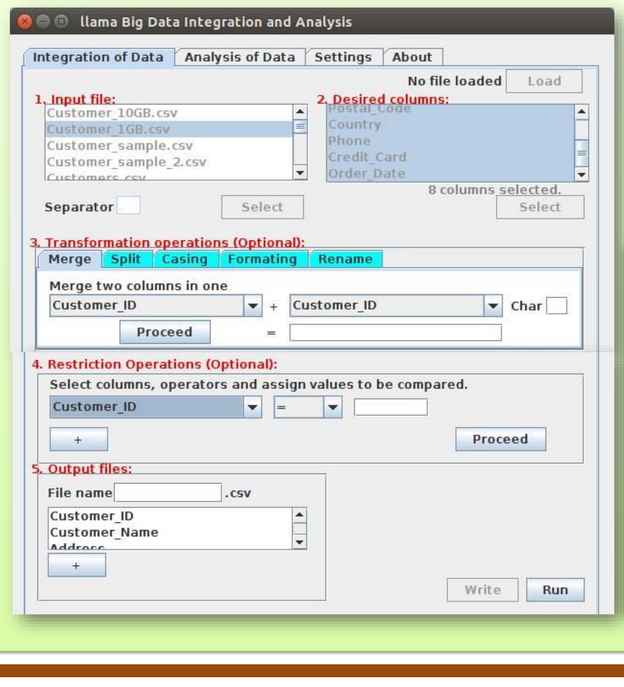
Implementation

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

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

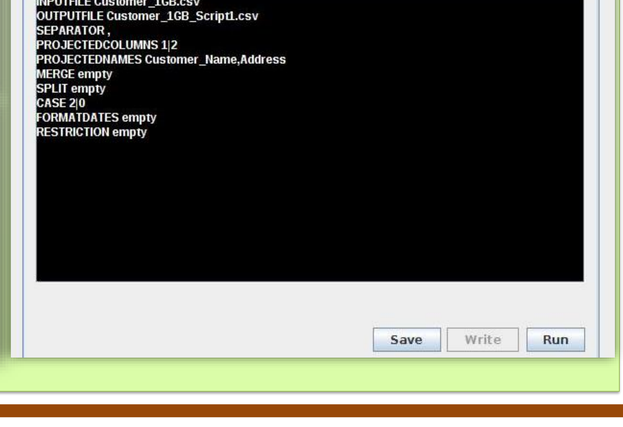
Data Integrator

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



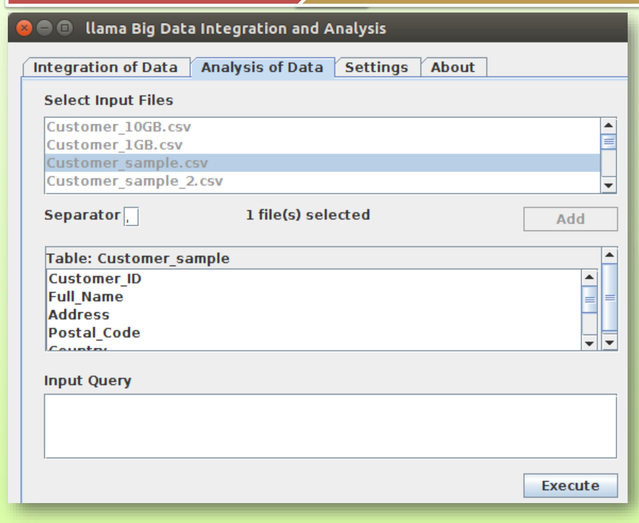
(*) alternative flow

Write or Load IJSL Script

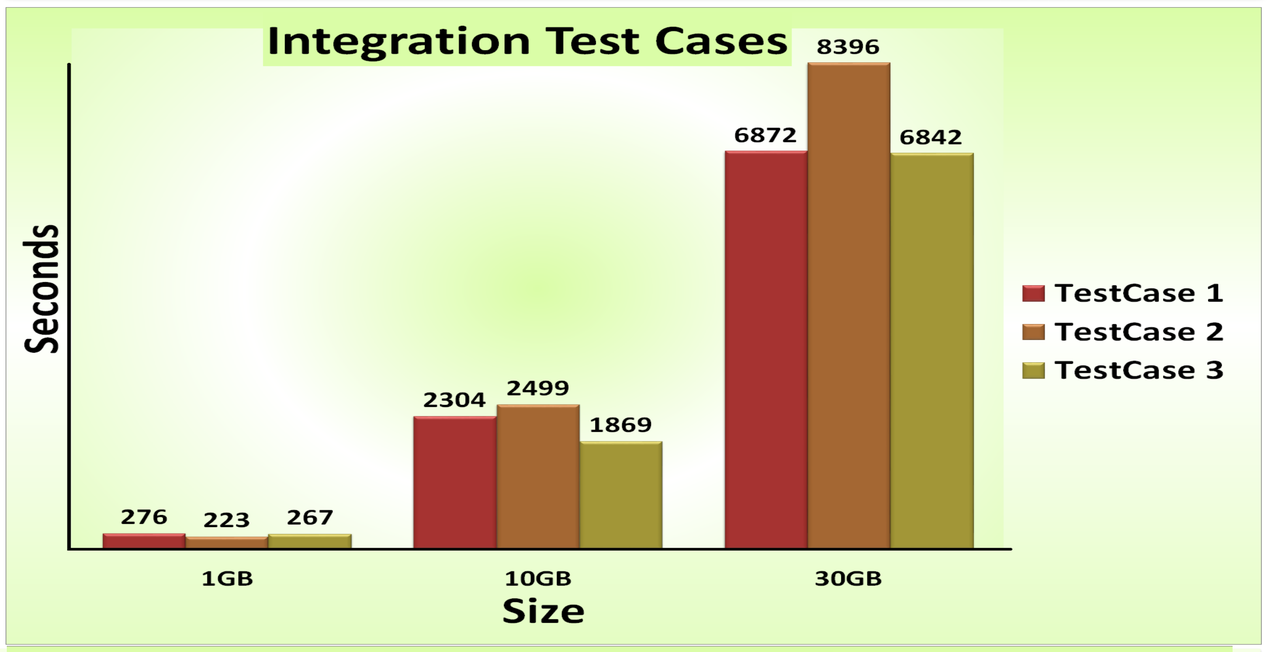


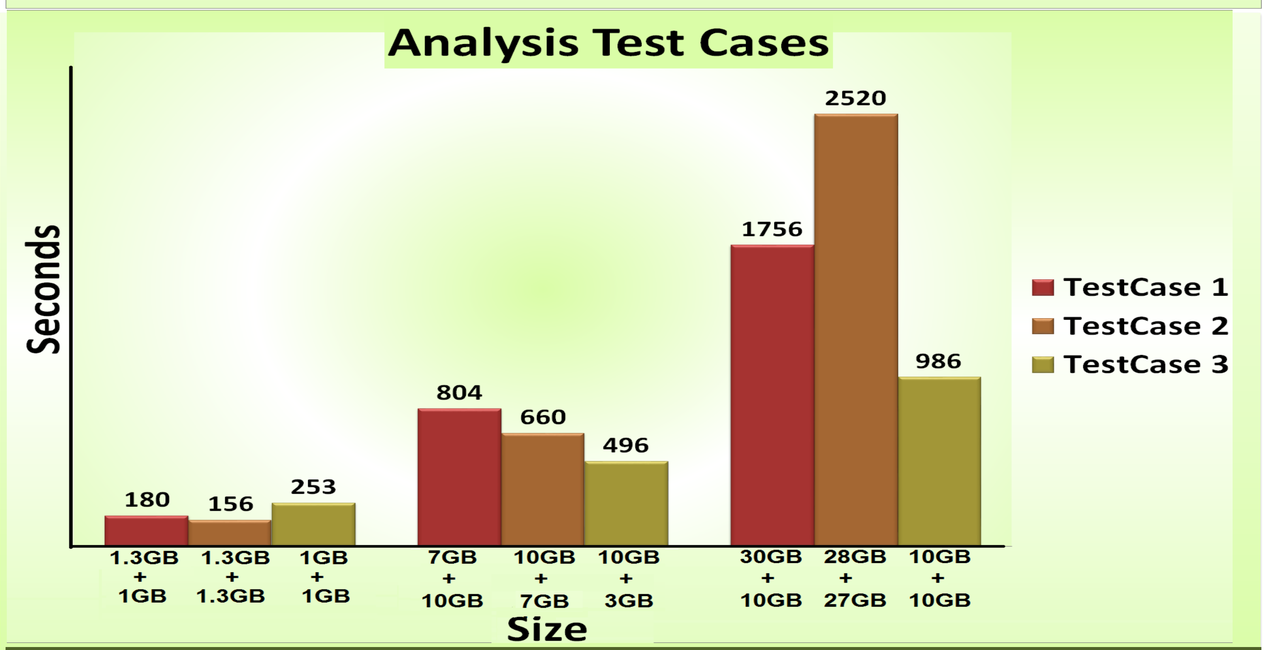
Data Analyser

Input File Selection → Writing Query → Visualization & Saving Results



Evaluations





Live Demo

Two virtual machines hosted on EIS02

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

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

Hadoop user:
User: hduser
Password: hduser

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

Remote Access Master:
Partner ID: 572120720
Password: 268vxh

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

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

EIS Enterprise Information Systems

universitätbonn