Big Data Integration and Analysis

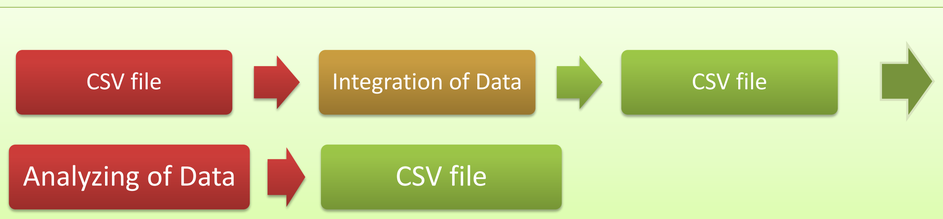
Authors:

Gaurav Kumar, Héctor Ugarte, Miguel Marmol, Tina Boroukhian

Summer Semester 2015

Definition of the Project

Goal: get the big data CSV file then Integration and Analyzing on it



Requirements: Ubuntu 14.4, Hadoop 2.6, Spark 1.4, Scala 2.1, Maven 3.3, Eclipse IDE, Java JDK 1.8
Input File: CSV File
Output File: CSV File

Integration & Analysis

Integration or **Cleaning Data** is the process of extracting, cleaning, formatting and loading data from a file, table, or database.

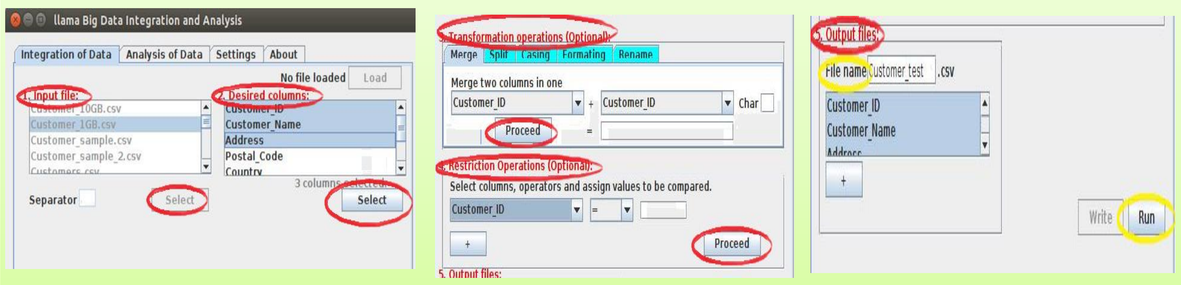
Analysing Data is a process of finding meaningful information and discovering hidden relationships in Big Data with the goal of supporting decision-making[1].

Apache Hadoop & Spark

Apache Hadoop is an open-source software framework written in Java for distributed storage and distributed processing of very large data sets on computer clusters built from commodity hardware.[3]
Apache Spark is an open source cluster computing framework. In contrast to Hadoop provides performance up to 100 faster for certain applications.[2]

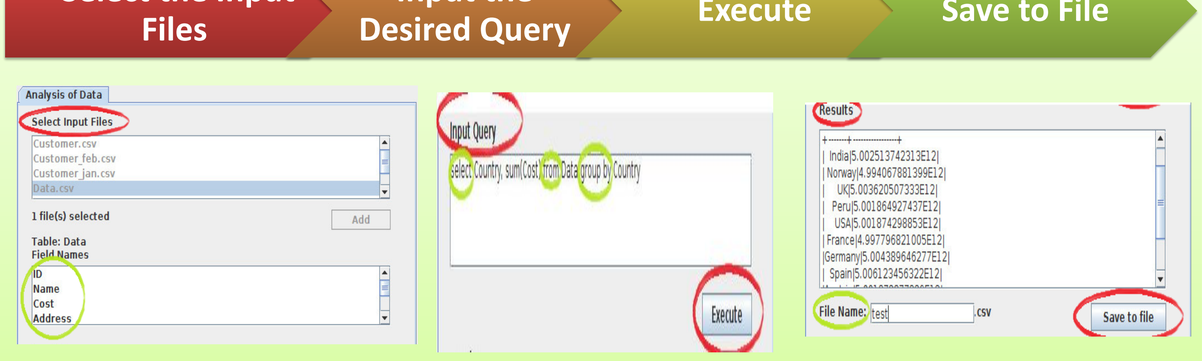
Integration of Data

Select the Input File → Select the Desired Columns → Select the Desired Operation → Select the Columns as Output → Run

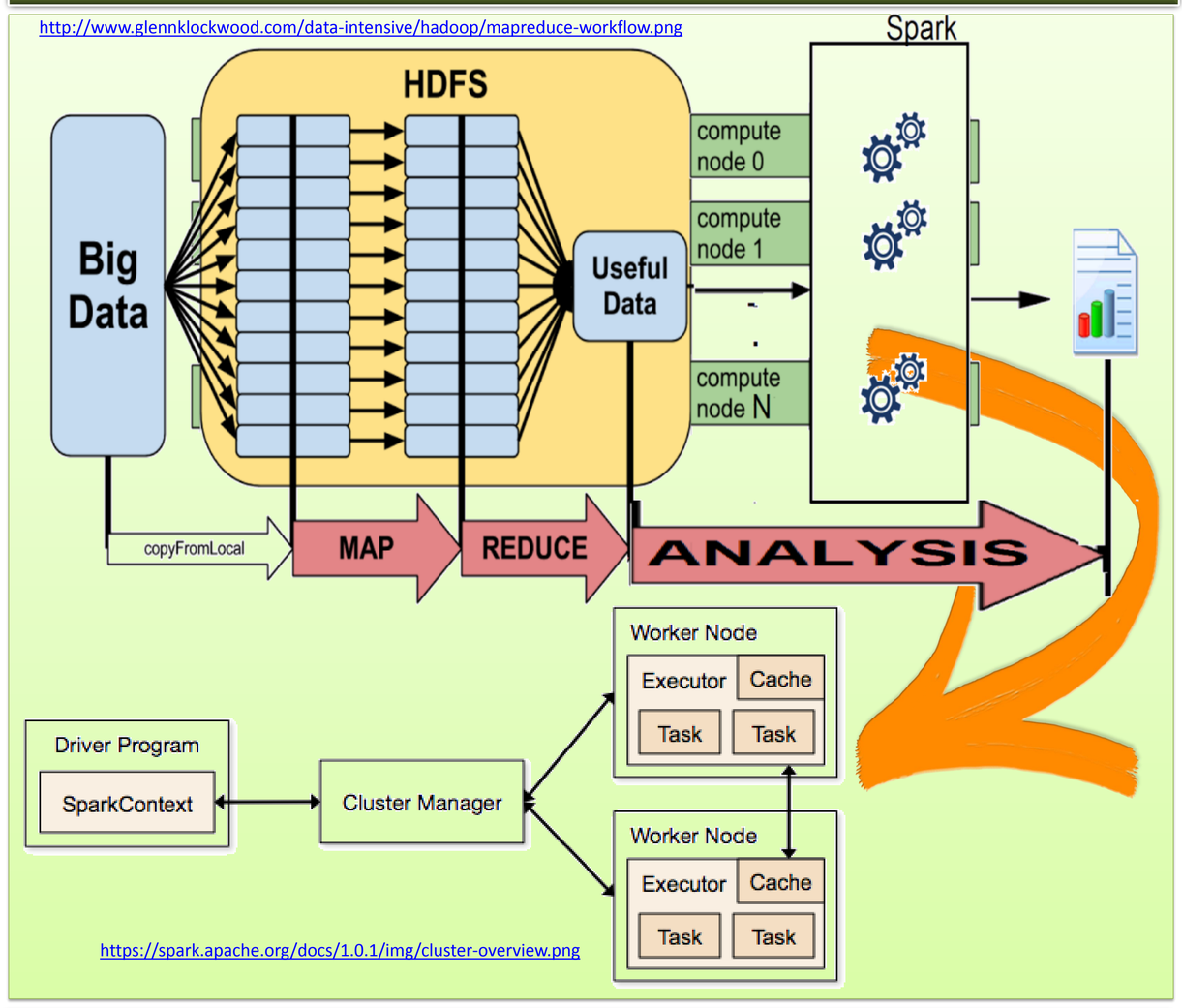


Analysis of Data

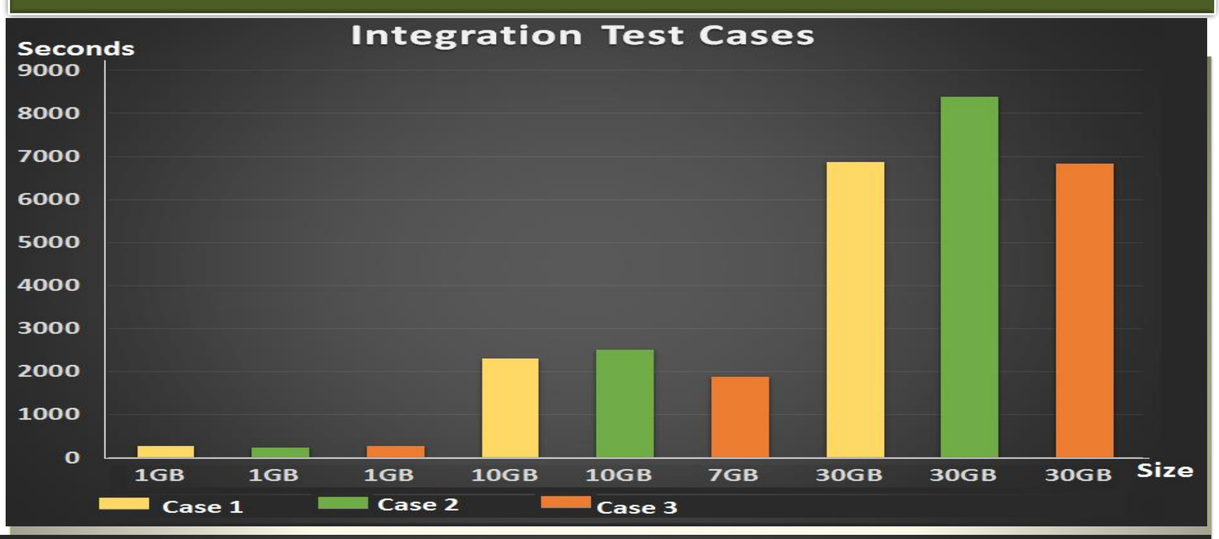
Select the Input Files → Input the Desired Query → Execute → Save to File

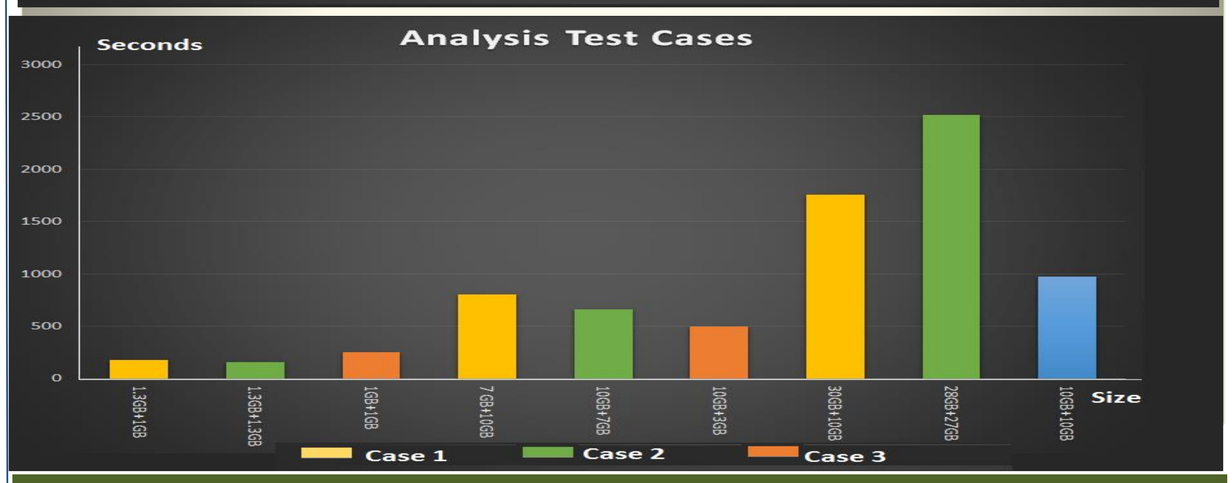


Architecture Diagram



Test Performance





Virtual Machine Description

Hostname: EIS03

Software installed:

Installed on:

- Ubuntu 14.04 LTS 64bit OS
- 1GB RAM
- Jdk 1.8
- Apache Hadoop 2.6
- Apache Spark 1.4

Local Access:
User: eis-user
Password: E1sbda2015

Remote Access:
Partner ID: 392368182
Password: E1sbda2015

Reference :
[1]http://www.oracle.com/technetwork/database/options/advanced-analytics/bigdataanalyticswpoaa-1930891.pdf
[2] https://en.wikipedia.org/wiki/Apache_Spark
[3] https://en.wikipedia.org/wiki/Apache_Hadoop

EIS Enterprise Information Systems

universitätbonn