Università degli Studi di Verona

Authorship Attribution

Big Data project report

Davide Bianchi VR424505
Matteo Danzi VR424987

November 19, 2019

Contents

1 Introduction

The project aim was to design a tool which could establish the authorship of a manuscript by using specific criteria described later. The used architecture is based on Hadoop, a distributed filesystem simulator, running in a docker container.

2 Background and System Description

A container is a standard unit of software that packages up code and all its dependencies so the application runs quickly and reliably from one computing environment to another. A Docker container image is an executable package of software that includes everything needed to run an application: code, runtime, system tools, system libraries and settings.

Docker containers images become containers when they run on Docker Engine. They isolate software from its environment and ensure that it works uniformly despite differences for instance between development and staging.

The Cloudera Docker image used in this project contains an Hadoop Distributed File System (HDFS) partition.

3 Project Workflow