Process Discovery using Big Data stack - Implementing the Alpha Algorithm with Map-Reduce

Design Specification Document

Martin Hashem, Xiangan Chen

May 7, 2019
RWTH Aachen

1 Introduction

Process mining is an approach to extract process models from event logs. Since the distributed nature of modern information systems, event logs are likely to be distributed among different physical machines. Map-Reduce is a scalable approach for efficient computations on distributed data. In this Python application we will present the main idea of a Map-Reduce implementation of the Alpha process mining algorithm, to take advantage of the scalability of the Map-Reduce approach.

To fulfill the shortcoming in the current process mining technology, this project aims to build a new python-based web service, that integrates the big data capabilities of the Hadoop system into the process mining framework pm4py.

This document is presented for our shareholders of the project.
Within our project we mostly apply well-established open source applications, which includes the following tools with version numbers:

- Python 3.6 as back end programming language
- pm4py 1.1.10 as process mining toolkit
- Flask 1.0.2 as web framework
- Hadoop 3.2.0 as big data processing and distributed computation framework
- Docker CE as deploy platform

This very document is one of these, which will be provided to bring insights to our software development cycle:

- Project initiation document
- Requirements analysis document
- Design specification document
- Software documentation

2 System Overview

The main goal of this project is to create a web application that provides calculations with the Alpha algorithm and can prepare the data using the Map-Reduce.

The web service is to be accessed by a standard browser and needs to provide the following functionalities:

- Upload event logs in either CSV or XES format into the Hadoop system
- Front end links to existing algorithms for further processing
- Do Map-Reduce computations to run the alpha algorithm
- Download the files onto local system

3 Design Considerations

3.1 Assumptions and Dependencies

The current situation in process mining eventlogs using the Alpha algorithm is very centralized. All eventlogs are pushed together into one computation unit, eating up a big amout of resources, both in used space and computation time. The optimization of this is the main task.

3.2 General Constraints

3.3 Goals and Guidelines

3.4 Development Methods

4 Architectural Strategies

5 System Architecture

6 Policies and Tactics

7 Detailed System Design

References

[de] docker enterprise. https://www.docker.com/.

[Fla] Flask. http://flask.pocoo.org/docs/1.0/.

[GA14] Joerg Evermann Ghazal Assadipour. Big data meets process mining: Implementing the alpha algorithm with map- reduce. *ResearchGate Conferance*, March 2014.

[had] APACHE hadoop. http://hadoop.apache.org/docs/stable/.

[Pg] RWTH Process and Data Science group. http://www.pads.rwth-aachen.de/.

[pmg] Fraunhofer FIT process mining group. https://www.fit.fraunhofer.de/en/fb/risk/process-mining.html.