Apache Spark API By Example

A Command Reference for Beginners

Matthias Langer, Zhen He
Department of Computer Science and Computer Engineering
La Trobe University
Bundoora, VIC 3086
Australia
m.langer@latrobe.edu.au, z.he@latrobe.edu.au

May 31, 2014

Contents

1 Preface

Spark is an advanced open-source cluster computing system that is capable of handling extremely large data sets. It was first published by **?** and its popularity has increased ever since. Due to its real-time properties and efficient usage of resources, Spark has become a very popular alternative to well established computational software for big data.

Spark is still actively being maintained and further developed by its original creators from UC Berkeley. Hence, this command reference and the associated, including the code-snippets and sample outputs outputs shown, should be considered as a overview of the status-quo of this amazing piece of software technology. Specifically, the API examples in this document are for **Spark version 0.9**. However, we do not expect the API to change much in future releases.

This document does not cover any installation or distribution related topics. For installation instructions, please refer to the Apache Spark website.