RHEEM

Zoi Kaoudi[§] Sebastian Kruse[◇*] Jorge-Arnulfo Quiané-Ruiz[§]

[§]Qatar Computing Research Institute (QCRI), HBKU
[◇]Hasso Plattner Institute (HPI)

{zkaoudi, jquianeruiz}@qf.org.qa {sebastian.kruse}@hpi.de

ABSTRACT

Bored of keep moving your app to the newest data processing platform to achieve high performance? Tired of dealing with a zoo of processing platforms to get the best performance for your analytic tasks? Then, this tutorial is for you!

Indeed, we are witnessing a plethora of innovative data processing platforms in the last few years. While this is generally great, leveraging these new technologies in practice bears quite some challenges, just to name a few, developers must: (i) find among the plethora of processing platforms the best one for their applications; (ii) migrate their applications to newer and faster platforms every now and then; and (iii) orchestrate different platforms so that applications leverage their individual benefits.

We address these issues with RHEEM, a system that enables big data analytics over multiple data processing platforms in a seamless manner. It provides a three-layer data processing abstraction with that applications can achieve both platform independence and interoperability across multiple platforms. With RHEEM, dRHEEMers (RHEEM developers) can focus on the logics of their applications. RHEEM, in turn, takes care of efficiently executing applications by choosing either a single or multiple processing platforms. To achieve this, it comes with a cross-platform optimizer that allows a single task to run faster over multiple platforms than on a single platform.

> "*... You may say I'm a* **d**RHEEM**er**. *But I'm not the only one. I hope someday you'll join us...*"
>
> – **John Lennon** –

Tutorial Plan

We plan a tutorial where participants are the main players and can interact with the speakers at any time. Tentatively, our tutorial will be as follows.

Introduction. In this first part of the tutorial, we will introduce the main concepts and rationale design behind RHEEM. This will allow participants to get familiar with the RHEEM paradigm.

Setting Up. After introducing participants to RHEEM, we will guide them through the installation and set up of RHEEM to get ready to code their first application.

Building Up your App. At this stage of the tutorial, we will show to users how one can easily develop and run an application on top of RHEEM. We will recap the main concepts in a hands-on exercise. In particular, we will see how to abstract and implement an application, how to support new functionalities by adding RHEEM operators, how to achieve higher performance by adding execution operators, and how to make the optimizer aware of the added operators.

Machine Learning. We will conclude this tutorial by summarizing all main concepts using ML4ALL, our machine learning system built on top of RHEEM and used by a big airline company from the middle-east. We will also leverage this application to walk participants through advanced features of RHEEM.

Speakers

Zoi Kaoudi is a Postdoctoral Researcher at the Qatar Computing Research Institute, HBKU. Her research interests include machine learning, big data analytics, and RDF data management. She has given several tutorials and lectures at SIGMOD'14, ICDE'13, Global Entrepreneurship Week'15, and Reasoning Web Summer School'14.

Sebastian Kruse is a Ph.D. candidate at the Hasso Plattner Institute in Germany and works as researcher in the Stratosphere project. His research interests include big data profiling, distributed data processing, and cross-platform data management. Sebastian has held several seminars on distributed data analytics.

Jorge-Arnulfo Quiané-Ruiz is a Scientist at the Qatar Computing Research Institute, HBKU. His research interests include cross-platform data management, big data analytics, and big data profiling. He has received an "Excellent Presentation Award" for his research talk at VLBD'14 and gave several tutorials and invited talks (at VLDB'12, MEDAL'16, EDBT Summer School'13, among others).

Hands-on Technology

We will use Spark, GraphChi, PostgreSQL, and a standalone Java engine to show the benefits of our system. RHEEM requires only Java 1.8 and Apache Maven 3.x. We thus assume participants have these two software tools installed

[*]Work done while interning at QCRI.

on their computer before the tutorial. Additionally, we will use a compute cluster to illustrate the full power of RHEEM. We will accompany our hands-on tutorial with supporting slides.

Rheem Website

RHEEM is publicly available on http://da.qcri.org/rheem/