# Thesis projects KPMG Big Data & Analytics team

## Introduction

KPMG offers data and analytics services to help organisations convert data from various sources into actionable insights that drive better decision-making. Often this conversion requires statistical analysis of large datasets, which is the area of expertise of the Big Data & Analytics team of KPMG The Netherlands ("big-data team").

The big-data team consists of data scientists, data engineers, data architects, and software developers, with backgrounds in physics and computer science. The team is led by Sander Klous, who is a partner at KPMG The Netherlands and a professor of *Big Data Ecosystems for Business and Society* at the University of Amsterdam. Several members of the team have a background in high-energy physics and gained experience in the processing and analysis of large datasets in projects at CERN.

Processing, storage, and distributed analysis of large datasets are enabled by KPMG's data-lakes solution. As a reference for the platform on which a data lake is built, the big-data team developed KAVE, which is based on Apache's Hadoop technology and contains the tools required for analysis of both small and large datasets. This platform is distributed as open source software, available on GitHub.

The open-source Eskapade framework is the main tool for data science and the rapid development of structured, production-ready data analyses. Eskapade is part of the KAVE platform, but it can also be used independently. The framework is built in Python, utilizing data-analysis tools like NumPy, Pandas, Apache Spark, Scikit-learn, TensorFlow and ROOT.

## Profile

- You are in the final year of your master's programme, typically in Computer Science, Mathematics, or a related field, with a focus on statistics, artificial intelligence, computational science, data science, data technology, or distributed computing.
- You have affinity with data science and (big-)data technology and are keen to learn about the software and methods applied in these fields.
- You would like to write your master's thesis on the topic of data science, data engineering, or distributed computing in the context of (big-)data analysis.

## Thesis projects

Several topics for thesis projects are available. You work on a project as an intern in the big-data team for a period of 6–9 months. If you are interested, please send an email to one of the contact persons listed below, or to Jeroen van Leerdam (VanLeerdam.Jeroen@kpmg.nl).

| Title | Category | Description |
|---|---|---|
| Statistical data quality monitoring<br><br>contact:<br>Max Baak<br>Baak.Max@kpmg.nl | Data science | Design and implementation of a method to read datasets of different given formats, convert to a unified format, and apply basic data-quality checks. Examples of input formats are CSV, JSON, Excel. The data may be converted to for example a Pandas or a Spark data frame. Quality checks involve checks of the data |

| | | |
|---|---|---|
| | | types of input variables, missing values, ranges of numeric variables, the format of string variables, the distributions of variables, and tracking these quantities over time.<br><br>The results of this project are to be integrated into Eskapade, a framework for reading, processing, and analysing data. The first reading and quality steps yield a dataset in a generic format, which is ready for further processing and analysis in the framework. A report with a summary of the input data should be generated, both as information to the user and as input for the next processing steps.<br><br>Scientific aim of this project is to monitor a range of incoming (e.g. financial) time-series data for a-priori unknown anomalies and trends and to quantify the significance, in order to predict in advance the sub-optimality or potential breakdown of (an) analysis model(s), such as financial risk models in the insurance domain, that require these data as input and now operate outside their usual operational bounds. |
| Simulation<br><br>contact:<br>Max Baak<br>Baak.Max@kpmg.nl | Data science | Development of an algorithm to generate realistic simulated data according to the multivariate distribution found in an input dataset. This would enable a data scientist to generate test datasets of arbitrary size, in which the variables follow the distributions of a prototype dataset. The simulation procedure can also be applied to evaluate systematic uncertainties associated to a given analysis procedure by varying or exaggerating features of the analysed dataset. In addition, the simulated samples can be used for oversampling for in classification problems with low signal to noise ratios.<br><br>The input distributions can be transferred to the generated data by for example creating a generic model built from Gaussian probability-density functions or by directly inverting the cumulative distribution functions of the input data. It is important that correlations between input variables remain intact. The procedure must also be able to handle variables of any type, e.g. simulate variables with string values. A possible extension is the simulation of data evolving over time.<br><br>Scientific aim of this project is to perform tests of existing complex analysis models (e.g. fraud detection models in the banking sector) for potential problems or inefficiencies, by simulating a) unit datasets with a well-predicted behaviour or b) realistic input datasets with controlled distortions. |
| Time Series Flattener<br><br>contact:<br>Max Baak<br>Baak.Max@kpmg.nl | Data science | Our clients often want actionable insights about possible future events. For example: their (end-of-year) revenue, customer's financial liquidity problems, customer churn, or mechanical failure. We investigate if it is possible to create a model which predicts these events with enough accuracy using as input both our client's data and publicly available data. For this to work, |

| | | |
|---|---|---|
| | | data often needs to be transformed into the proper form, and features need to be generated from the transformed data.<br><br>The software goal of this project is to develop a generic method to transform multivariate time series data, possibly originating from many sources (e.g. many components in a railroad network), combined with static dimensional data into a (machine-learning) training dataset. Features must be extracted from shifting time windows of variable lengths. Optionally, labels must be extracted from successive time windows of variable lengths. The windows needs to be tuned to give optimal performance.<br><br>The scientific aim is to give an overview of currently available methods which can (partially) do the same, and study how well they work in comparison with our novel method. A statistically quantitative indicator must be created for comparison purposes. We will test this by applying the algorithms to a) financial transaction data or b) Wi-Fi sensor data used for location aware services. |
| Software architectures for parallel machine learning<br><br>contact:<br>Riccardo Vincelli<br>Vincelli.Riccardo@kpmg.nl | Data engineering | As the data loads get bigger, reducing the time needed to figure out the right model to go with becomes crucial. Traditionally large scale parallel deployments for machine learning solutions are based on specialized hardware supports, usually GPUs.<br><br>The idea is to enable massively-parallel big-data-science pipelines. The result of the work should be integrated into Eskapade, the decision-engine framework developed by the big-data team. The investigation and work aim at making this synergy concrete on Eskapade, relying on its integration with the Hadoop framework, in particular Spark and, to a deeper extent, YARN.<br><br>We will address technological questions like:<br>• Is it possible to make YARN GPU-aware, what is the state of the art in Hadoop here?<br>• Alternatively, would a micro-services-only architecture be meaningful in this case, for example being Docker GPU-aware?<br>Part of the answer will of course consist in modifications and evolutions applied to the source code of Eskapade, a working solution. |
| Micro-service architectures: evolution of existing solutions<br><br>contact:<br>Riccardo Vincelli<br>Vincelli.Riccardo@kpmg.nl | Distributed computing | The architectural approach of micro-services is an interesting and appealing way to design high-scale software solutions. Drawing from traditional concepts such as UNIX network-coordinated processes as well as REST API interfaces, but enabled by the powerful advancements in virtualization software, micro-services architectures have become central in the production agendas of both start-ups and enterprises. |

| | | |
|---|---|---|
| | | This work aims at putting the micro-services architecture into context by applying it to a concrete software project. The ideal development is as follows:<br>• theoretical and historical background, with emphasis on different architectures (actors, containers, lambdas, etc.)<br>• redesign of an existing production solution into the micro-services architectural pattern<br>• implementation in one or possibly multiple programming languages |
| Security topics in REST API management<br><br>contact:<br>Riccardo Vincelli<br>Vincelli.Riccardo@kpmg.nl | Distributed computing | Modern software solutions are adopting REST APIs as the preferred way to support user interaction. In the evolving landscape of smart solutioning, such as ad-hoc Lambda micro services, security has become a central concern, since the solution and the data it operates on are more often than not distributed in the cloud.<br><br>This thesis is aiming at validating an existing software solution in terms of security standards. Whereas the starting point can be the documentation and material provided by OWASP, the student can also refer to other sources as well as case studies from the industry. The chosen solution will be checked against standards and cases for security as above, but also corrected and enhanced as needed to match them. |
| An Open Source solution for data lakes<br><br>contact:<br>Riccardo Vincelli<br>Vincelli.Riccardo@kpmg.nl | Distributed computing | The data lake is the core of the big-data ecosystem for the enterprise. Most data-lake solutions are based on the Hadoop framework. KPMG offers its own Hadoop distribution, KAVE. KAVE has been successfully adopted by clients worldwide, and is also an advanced software solution for data lakes.<br><br>The aim of this thesis work is to understand what the state of the art in technology is for data-lake solutions and improve the KAVE as a valid Lakes all-in-one solution. In concrete terms, you will work on implementing novel data processing functionality and services into the KAVE with support by the KAVE technical leads and internal stakeholders. |