European Commission

DEDS

European Joint Doctorate in
Data Engineering for Data Science (DEDS)

Doctoral Project Plan[1]
Thesis Title
First Name Last Name

Section 1. Project Summary

The term Big Data refers to data sets that are too large or complex to be dealt with by traditional data processing software. In particular, Big Data captures to 4 dimensions: Velocity, Variety, Veracity, and Volume. This large quantity of data available nowadays has a lot of statistical and business value, therefore their analysis is of core importance for business decision making.

Data Integration is the set of processes to gather and bridge data from heterogeneous sources together in order to have a unified view. The premise of data integration is to make data more freely available and easier to consume and process by systems and users. Research on Data Integration started more than 50 years ago [6, 9] but as we entered the Big Data era, new challenges arose, which include scaling [4] data integration while guaranteeing the privacy [10] of the individuals involved in the datasets. Over the last decade new techniques have been applied for improving computational performance, consisting of the use of parallelizing the computation by using big data processing platforms [4], or algorithmically, namely using summarization techniques [2], used for approximate fast and approximate querying, improving the performance of Machine Learning processes [7, 1, 8] as well as in some data integration scenarios.

Analyzing Big Data may involve accessing sensitive information, especially in the health sector. For these reasons, in Data Integration, techniques like Secure Multi-Party Computation and Differential Privacy [5] have been used. One big concern in private Data Integration is that it is difficult to guarantee privacy, accuracy, and good performance at the same time.

Machine Learning and Data Integration have really close relationship [3]. In particular, it is possible to leverage the first to improve the performance of the second and vice-versa. An evolving branch of Machine Learning is Federated Learning, which consists in building a Machine Learning model in a federated setting (when the data is distributed across edge devices) and model in a collaborative way, without moving the data to a central server.

This Ph.D. aims to explore summarization techniques and Federated Learning for scaling Data Integration tasks while guaranteeing privacy and achieving good performance.

TPR: An updated version of the summary, concretising key motivation, significance, methodology, and expected outcome of the doctorate study.
RPR: An updated version of the summary, refining key motivation, significance, methodology, and reporting the hitherto outcome of the doctorate study.

Section 2. Scientific Content of the Doctorate Project

1. The background for the project problem should be described (maximum 300 words).

[1] Choose the appropriate heading. Based on the PhD Study Plan of Aalborg University, available at `http://www.phd.teknat.aau.dk/intranet/phd-study-plan`.

2. An introduction stating the state of the art for the doctorate project. The introduction should include key references listed under Section 7. Typically, at least 10-15 references to peer-reviewed scientific material are expected. In case it is necessary to refer to non-peer-reviewed material then use a footnote (or parenthesis) to provide information to the source. Explain the relevance of the present doctorate project so the scientific contribution will be evident – i.e., explain how the project advances current state-of-the-art. Scientific challenges should be clearly defined – do not mistake this for technological challenges.

TPR: The state of the art for the doctorate project must be updated including use of the most essential references (list references under Section 7).
RPR: The state of the art for the doctorate project must be updated including use of the most essential references (list references under Section 7). The relevance of the project results achieved so far should be reported and how they advance the current state-of-the-art.

3. Statement of the project's objectives. This could be formulated as a hypothesis and/or research questions if applicable.

TPR: Updated statement of the project's objectives followed by a formulation of the specific problem(s) that is(are) to be addressed in the study. The problem(s) could be stated as one (or more) scientific hypothesis, if relevant, that is (are) to be examined.
RPR: The final statement of the project's objectives must be specified, followed by a formulation of the specific problem(s) addressed in the study.

4. Key methods. Coverage of the methodological needs, identification of means of meeting these needs, and the methodological design. The coverage should include techniques for evaluating or assessing the outcome of the project (e.g. empirical studies and/or theoretical studies).

TPR: Update the preliminary key methods, planned for the doctorate project.
RPR: Report the final key methods used for the doctorate project.

5. Potential significance and application(s) of the project's expected outcome, possibly including methodological contributions.

TPR: Experiences and results obtained so far in the project followed by the expected outcome of the entire doctorate project. What is the potential significance of this expected outcome, possibly including methodological contributions.
RPR: Report on the experiences and results obtained in the doctorate project, as well as the expected ones.

6. Work and time plans including measurable milestones (project milestones and deadlines for expected publications for each 6-month period, or finer). A practice is recommended where results are documented and submitted for publication in peer-reviewed outlets throughout the project.
The planned timing for the stays at the host institution should be given. In addition to this, the plans for mobility to a partner organisation (secondment) should be stated.

TPR: An updated time schedule for the entire project must be included. It is recommended that a number of subproject activities are identified that can be associated with milestones, so that there are milestones (at least) each six months during the project. Remember to allocate time for preparing scientific publications (conference papers, journal paper, etc.). Indicate deadlines for the expected publications. These milestones will allow the doctorate candidate and co-supervisor(s) to assess the status of the project each six months and to

revise the plan if needed. The specific activities described in the time plan must be of such detail that it is clear what should be carried out.
RPR: Report on the fulfilment of the planned activities, and concrete time schedule until the finalising of the thesis.

7. Outline of the content of the thesis.

TPR: This description could be organised by means of an overall table of contents.
RPR: The final, refined outline of the thesis.

8. Outline the publication strategy for the project. Tentative titles (or expected subjects) on papers, including preliminary authors list (indicate who has the primary responsibility for the publication). Three international peer-reviewed publications should be planned, at least.

TPR: For each publication, the following should be indicated or estimated: working title, co-authors, length in pages, outlet (e.g. a named conference or journal), and approximate time of submission. Indicate who has the primary responsibility for the publication.
RPR: The list of publications, published and/or accepted as part of the doctorate project should be reported, together with the ongoing and planned ones.

Section 3. Co-supervisors/Candidate Co-operation Agreements

The project will be carried out in three years during which the PhD student will stay in one research institution and one university. During the first and the third year, the candidate will work in Athena Research Center (ARC) under the supervision of Prof. Minos Garofalakis (ARC). During the second year, the program will take place in Universitat Politecnica de Catalunya (UPC) under the supervision of Prof. Oscar Romero (UPC). The project will be a joint work of all parties, hence close co-operation is expected in the following way.

The progress of the project will be validated through frequent meetings between the candidate and his supervisors. The candidate will meet on a weekly basis with his home supervisor and one or two times per month with his host supervisor (the opposite when he will be hosted at UPC). Following typical business practice, the expectations and tasks planned for each meeting will be clearly communicated in advance, with a reasonable notice, both from the supervisors to the candidate and vice-versa. Standard tools of the trade will be used to boost collaboration, such as a shared repository for documents and code artifacts (e.g., Mendeley Library, Github), communication platforms (e.g. Skype, Teams).

Section 4. Proposed Education and Training Programme

The DEDS education and training programme is composed of several activities.

- **Research**, where doctoral candidates work on a novel research problem guided by two supervisors who will advise them to gradually become independent researchers.
- **Research-specific courses**, aimed at providing doctoral candidates with focused state-of-the art technical skills pertaining to their research topic.
- **Innovation and entrepreneurship courses**, aimed at complementing the scientific training of doctoral candidates with business-related aspects such as entrepreneurship, intellectual property rights, etc.
- **Methodological and communication courses**, aimed at introducing the necessary research methods and communication skills.
- **Language courses**, aimed at introducing the local language at each partner university.

- **Summer and winter schools**, wherein candidates will obtain feedback about their research from invited researchers and practitioners, as well as get international contacts in both academia and industry.
- **Tutoring**, whereby candidates will be involved in teaching activities (e.g., supervising student projects and delivering exercises) while being coached by their supervisors or other experienced staff.
- **Knowledge dissemination and participation to scientific events**, aimed at allowing doctoral candidates to present and confront their findings, thereby familiarising themselves with essential practices such as peer-review and public debating.
- **External cooperation and secondments**, aimed at ensuring that the candidate participate actively in another research environment outside his/her home and host universities. These activities are realised typically with DEDS partner organisations.

Please detail in the following subsections your personalised education and training programme, taking into account your previous background and future career prospects. This programme must be approved from both co-supervisors.

Activities adding at least 30 ECTS credits must be outlined. A tabular listing of all activities performed or to be performed during the doctorate project is to be included. Group the activities according to the categories specified above. For each activity, the title, time, location, organiser, and ECTS credits must be included together with an indication of whether the activity has been completed. Please use this table:

Activity	Place/Organised by	ECTS	General/Project course	Status

TPR: The table of activities must be updated with planned and hitherto completed activities.
RPR: The contents and the extent of the completed activities must be reported. It is expected that all training activities have been finalised in order to devote the last year of the project for finalising the Doctoral Dissertation.

Section 4.1. Planned Courses

Courses adding at least 20 ECTS credits must be outlined.

Only courses at doctorate level are approved. If a course at master level is deemed to be highly relevant for the doctorate project, the co-supervisors can establish a study group on the topic, which includes the master course and additional reading/discussion to bring it up to doctorate level. A written report on participation in a study group must be completed to get course credit. To ensure the scientific level, the study circle must be headed by a member of the scientific staff, who is Professor or Associate Professor (senior scientist level). A 2-3 ECTS study circle organised by the co-supervisors on the state of the art in the research field of the doctorate study is recommended.

TPR: The course table should be updated with more specific information for the completed courses, as well as with the rest of planned courses for the rest of the doctorate project.
RPR: The table must be updated, reporting the complete set of courses that are completed in the doctorate project.

Section 4.2. Knowledge Dissemination and Participation to Scientific Events

Detail the plan for dissemination of knowledge and findings from the project.

This could for instance be:

- Poster presentations at conferences/seminars.
- Presentations at conferences/seminars
- Newspaper articles or other popular presentations
- Teaching (lecturing and project supervision)

Each participation to scientific events must be accompanied by a written report by the doctorate candidate that relates the specific activity to the doctorate project. This report must be of general value for the project. Activities that relate to workshop and conference participation must not exceed 6 ECTS credits.

TPR: Updated plan for dissemination of knowledge and findings from the doctorate project other than those listed in Section 2 must be specified.
RPR: The final realisation of the knowledge and findings dissemination from the doctorate project other than those listed in Section 2 must be reported.

Section 4.3. External Co-operation

The doctorate candidate will spend time studying both in Greece ‘ and Spain. Furthermore, a secondment of three months will take place, where the candidate will join Spring Techno, where he will work on of a complex Federated Learning scenario with real data. During the following three years, all ESRs will meet in four different winter/summer schools to present their work, receive feedback, exchange ideas, and get exposed to new challenges. During these schools, candidates will have the opportunity to get in touch with academic and non-academic partners, presenting them their findings, reflecting on new opportunities, and opening the way for further collaboration. Finally, the candidate may co-operate with external researchers or research teams, in case that his work can be combined or merged with similar works of others.

Section 5. Agreements on Immaterial Rights to Patents

Patents and immaterial rights will be handled according to general rules applied by Athena Research Center, National and Kapodistrian University of Athens, and Universitat Politecnica de Catalunya.

Section 6. Financing Budget

This project is one of the 15 ESRs of Data Engineering for Data Science PhD programme, which is funded by the European Union’s Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 955895. The funding covers expenses related to the successful completion of the project, such as work equipment, research experiments, training activities and others that are relevant to the programme.

Section 7. Career Development Plan

In this section, the candidate’s career plan and development are described. When the thesis is handed in (M4), this section is revisited in a self-contained document called “Career Development Plan” (CDP) to be signed by the candidate and the supervisors.

Section 7.1. Long-Term Career Objectives

DPP: Describe long-term career goals (over 5 years) and how to become able to reach those goals.
TPR, RPR, CDP: Update as needed if the career plans have evolved.

Section 7.2. Objectives Covered in Project

Describe which development objectives will be/have been achieved in the project with respect to

1. Research skills and techniques
2. Research management and co-operation
3. Communication skills
4. Other professional training
5. Networking activities and opportunities
6. Other activities with professional relevance

CDP: Include also published and accepted papers as well as completed course activities such that the document is self-contained.

Section 8. References

References

[1] Jesus Antonanzas, Marta Arias, and Albert Bifet. "Sketches for Time-Dependent Machine Learning". In: 1 (2021), pp. 1–9. arXiv: 2108.11923. URL: http://arxiv.org/abs/2108.11923.

[2] Graham Cormode et al. "Synopses for massive data: Samples, histograms, wavelets, sketches". In: *Foundations and Trends in Databases* 4.1-3 (2011), pp. 1–294. ISSN: 19317883. DOI: 10.1561/1900000004.

[3] Xin Luna Dong and Theodoros Rekatsinas. "Data integration and machine learning: A natural synergy". In: *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2019), pp. 3193–3194. DOI: 10.1145/3292500.3332296.

[4] Xin Luna Dong and Divesh Srivastava. "Big data integration". In: *Proceedings - International Conference on Data Engineering* September (2013), pp. 1245–1248. ISSN: 10844627. DOI: 10.1109/ICDE.2013.6544914.

[5] Cynthia Dwork and Aaron Roth. "The algorithmic foundations of differential privacy". In: *Foundations and Trends in Theoretical Computer Science* 9.3-4 (2013), pp. 211–487. ISSN: 15513068. DOI: 10.1561/0400000042.

[6] Ivan P. Fellegi and Alan B. Sunter. "A Theory for Record Linkage". In: *Journal of the American Statistical Association* 64.328 (1969), pp. 1183–1210. DOI: 10.1080/01621459.1969.10501049.

[7] Rémi Gribonval et al. "Sketching Datasets for Large-Scale Learning (long version)". In: (2020), pp. 1–35. arXiv: 2008.01839. URL: http://arxiv.org/abs/2008.01839.

[8] Jiawei Jiang et al. "SketchML: Accelerating distributed machine learning with data sketches". In: *Proceedings of the ACM SIGMOD International Conference on Management of Data* (2018), pp. 1269–1284. ISSN: 07308078. DOI: 10.1145/3183713.3196894.

[9] Ramez El-Masri and Gio Wiederhold. “Data Model Integration Using the Structural Model”. In: *Proceedings of the 1979 ACM SIGMOD International Conference on Management of Data*. SIGMOD ’79. Boston, Massachusetts: Association for Computing Machinery, 1979, 191â202. ISBN: 089791001X. DOI: 10.1145/582095.582127. URL: https://doi.org/10.1145/582095.582127.

[10] Shui Yu. “Big privacy: Challenges and opportunities of privacy study in the age of big data”. In: *IEEE access* 4 (2016), pp. 2751–2763.

European Commission

DEDS

European Joint Doctorate in Data Engineering for Data Science (DEDS) Doctorate Project Plan[2] Thesis Title First Name Last Name

This page must be completed and sent together with the project plan/report in a pdf file to the chair of the Candidate Progress Committee.

Project title: .....
Name of doctorate candidate: .....
Email: .....
Supervisor: .....
Home University: .....
Co-supervisor: .....
Host University: .....
Secondment supervisor: .....
Partner organisation: .....
Date of enrolment: .....
Expected date of completion: .....

Signatures

The Doctorate Candidate

.....
Date:

The Supervisor from the Home University

Professor .....
Date:

The Supervisor from the Host University

Professor .....
Date:

The Secondment Supervisor

.....
Date:

[2]Choose the appropriate heading among the three

The Chair of the Candidate Progress Committee

Professor .....

Date: