BDMA JOINT PROJECT

Introduction

The BDMA joint project follows a project-based learning paradigm. The main objective is learning by doing and complementing the courses lectures and labs.

The project simulates a real start-up project: first, you must come up with an innovative data-oriented business idea. During this stage it is important you understand how to assess the viability of a business project. The tools and techniques you learnt at VBP should be put in practice in that part. Note that, from a big data and data science point of view, there is a main challenge you should cover here:

- Describe the added value of your project in a precise and concise manner.

Even if the course is oriented to communicate with business people, note that this is an important skill in general (to promote your project within an organisation or to convince someone to start a project in whatever scenario).

Next, you must develop a prototype showing the feasibility of your business idea from an IT point of view. Three courses are involved in this part: CC, BDM and SDM.

- In CC you should explore what infrastructure and services (potentially from Cloud providers) are of interest for your project.
- In BDM you must develop a full-fledge functional architecture and data flows, reaching up to the tool selection.
- Finally, your project should exploit and benefit from graph-based formalisms as taught in SDM.

Each course evaluates certain aspects of the project. In this document, we will define the aspects covered in SDM.

Statement

Include graph-based solutions (either property graphs or knowledge graphs) into your project. The two main alternatives to benefit from these formalisms are as follows:

1. **Graph-based analytics**: Create a data view (or a data repository) in the form of a graph and use graph analytics. For example, you may instantiate a lambda architecture for BDM and create a view in the serving layer in the form of a graph database. The two main tasks here would be to design and populate the graph and later use graph analytics.
2. **Create a data catalog to facilitate data integration**: Use knowledge graphs for conducting advanced data integration tasks. For example, this is a good option for a project with many sources. In these cases, you can automate, to some extent, the integration of data coming from different sources leveraging on a data catalog.

The overall objective is to experiment with these technologies in real projects.

Tasks to conduct

[M1] Decide how you are going to use graph-based solutions in your project. Then, **create a brief purpose statement** mentioning how the project benefits from your proposal.

[M2] Decide and **justify the most appropriate graph family** for the problem. Either property graphs or knowledge graphs (in this case, you must also choose the appropriate language).

[M3] Design the graph. If possible, tailor the graph schema with a meta-model definition. Otherwise, describe the schema and provide some instance. *Design the graph using the bubbles and arrows metaphor used during the course*.

[M4] Design the flows to populate the graph. Basically, you must ask the following question: from what sources is the graph populated? As we have discussed during the course, populating the instances should be a (semi-)automatic process.

[M5] Explain the processes exploiting the graph. Describe the processes used in your project to exploit the graph. This explanation should be aligned with M1.

- If the graph is meant to be a data catalog, how the catalog is used to automate data integration tasks. *You may want to use BPMN to precisely describe these processes*.
- If it is a general-purpose graph, describe the graph analytics and their purpose. In this case, you must use the usual steps to describe data analysis: *preparation, creation of the model (including its parametrization) and validation tasks*. Include a discussion about the added value of such analysis in terms of the project. Note that you can use any algorithm (in the literature you can find a plethora of graph-based algorithms), even data mining (DM) or machine learning (ML) algorithms on top of graphs (in this case, be sure it makes sense to start from a graph to conduct DM or ML).

[M6] If the processes in M5 and/or M6 generate metadata (e.g., for reinforced learning), please, explain **how metadata is generated, stored and (re-)used**. *You may want to use BPMN to precisely describe these processes*.

[M7] Implement a proof of concept (PoC). Choose the appropriate tools and be sure you can execute and end-to-end example to showcase the previous points. Here, you may need a graph database / triplestore, a program potentially using graph frameworks and a query engine (that will depend on the graph family chosen). Clearly describe the PoC setting and justify the decisions of the tools chosen.

Deliverables

First, use the team creator event to register your joint project into SDM.

Once done, by the deadline stated in this event, one person of the group must upload a document giving answers to M1-M7. There is no need for very large documents. Simply, be precise and concise. Therefore, **a maximum length of 10 content pages per group is set**.

The document must include a link to the github or project page you are using (if it is a private repository, please, create a user and state login / passwords credentials in this document).

Timeline

To help you at the beginning of the project, we will set one online sessions where we will present the main project ideas and you can ask questions. See the course schedule and look for the “project” session. In this session you should work with your group and clarify questions with the lecturer. Later, as you progress with the project, each group, if needed, can schedule an individual 1-1 session with the lecturer to clarify their doubts.

Finally, there will be a joint defense (more details will be available further on in the Big Data Management and Analytics meta-course).

Evaluation criteria

The document will be evaluated according to the following criteria:

Conciseness

- Report: max 10 pages
- Appendix 1 (optional): tool selection matrices (excluding any discussion): max 2 pages
- Appendix 2 (optional): flow diagrams (excluding any discussion): max 2 pages

Understandability

You provide enough details as to assess your solution.

Coherence

The solution is well-integrated in the project, makes sense as a whole.

Soundness

There are no contradictions about the choices made and the inherent advantages of the underlying theory chosen.

Note: the document will account for 60% of the mark. The joint defense accounts for the 40%.