Big Data Management (BDMA & MIRI Masters)
Project: Big Data architectures

Sergi Nadal, Alberto Abelló, Oscar Romero

1 Introduction and rationale

As you may have realised, Big Data is still in a very preliminary condition. This means that standards and mature frameworks are completely missing. As result, the way to approach a Big Data project is completely different to traditional software projects. Big Data is about specific solutions for specific problems (typically characterized with the three V's: Volume and / or Velocity and / or Variety). Thus, you cannot expect to find golden rules (there is no universal solution even if some companies claim for it; first lesson: do not fall in marketing traps) or frameworks that are going to hide most of the difficulties (again, be very careful with what is claimed and better focus on what a given solution can really do). However, it is a matter of time that complex frameworks hiding most of these complexities appear but, until now, a Big Data specialist needs to master many different aspects and do really know the technology s/he is using.

Big Data tools are far away from being mature. The only way to know how they are going to behave is by understanding how they work internally. That is why a strong knowledge on the data management and analytics is needed. When starting a Big Data project you need to wonder the following (before coming up with a solution):

- **[A]** What are your data sources? Where is data arriving or when is going to be ingested? At which pace / velocity? What is the expected volume? Do you have a serious heterogeneity problem (i.e., variety)? It is important to know the availability of the sources, how to connect to them and the arrival rate of data. In short, you need a comprehensive characterization of your data.

- **[B]** How are you going to model your data and store it? Remember the golden rule of data management, the way you store your data tells you what you can do with it. Thus, you need to know how you are going to query your data and choose the right architecture by considering the internal characteristics of the Big Data tools considered. Without this, your Big Data solution most probably will suffer from performance issues sooner than later, or if not, most probably you did not have a Big Data problem in hands.

- **[C]** What algorithms are you going to use to exploit your data? Your queries may range from typical reports or one-record queries, to range queries, large sequential reads and complex analysis by using Machine Learning / Data Mining. Note that in the later it is important to consider the cleaning and preparation transformations you need to perform on data and how to efficiently compute such algorithms in data most probably distributed in a cluster. It is also extremely important to model your data such that:
  - you exploit sequential reads as much as possible,
  - benefit from indexing and pre-fetching,
  - and maximize the effective read ratio (e.g., by using vertical fragmentation).

For all these reasons, it is impossible to map one Big Data project to another unless they share the same aspects (similar queries, data, constraints, etc.). Most probably, every Big Data project will need its own solution. One may say though: what about experience? Is it not a plus here? Of course, you will have many intuitions and experiences that can be applied from one project to another, but you need to carefully characterize your problem and know the internals of the tools in order to get there and be sure that a solution can be mapped. Trying to apply solutions from other projects without considering the specificities of the current project is a typical mistake in Big Data. For all these reasons, this project aims at gaining experience and realise by yourself about all these issues.

2 Statement - Implement a business idea

In this project you a required to propose a technical solution for a business idea you develop[1]. The project has four main tasks:

1. Identify a business need (and the data sources),
2. propose a technical solution for the business need,
3. describe the data flows involved,
4. develop a Proof of Concept (PoC) showing the feasibility of your solution.

3 Tasks to conduct

3.1 Identify a business need and its data sources

Here, you are asked to think about a business idea that requires a Big Data solution. The objective of this first phase is to realise about the impact of your solution in a business. Special attention should be paid on the potential data sources where to extract data from. Describe each of the sources involved, what is their role within the project and the specific characteristics of each.

3.2 Propose a technical solution

Now it is time to characterize your problem and as consequence, justify what is the best architecture. Here, you are asked mainly to consider how are you going to ingest and store data. Note this includes deciding what data to store to use and how to model them. Identify functional components that will compose your architecture, later identify what tools could be used to implement such components.

3.3 Describe your data flows

Once you have a clear idea of your architecture and tools, it is time to describe your data flows. Identify different analytical scenarios and, starting from the sources, describe the input and output data for each of the components in the architecture, also describe what transformations are applied. Additionally provide performance and quality metrics of the described data flows, present figures and their analysis.

3.4 Implement a proof of concept

Once your proposal is ready, you are asked to develop a proof of concept (PoC) showing the feasibility of your proposal. There will be a final presentation where you need to motivate and explain your business idea as well as the technological solution designed and show the PoC. This part will be assessed with the technical deliverables and the final presentation showcasing your PoC.

[1] If you are enroled in VPB, you are expected to work on the business idea developed in that course. Thus, you are expected to work with the same peers as in VBP.

4 What is expected?

A single deliverable is expected per team. Clearly state the team identifier and the names of all members. Do not modify the margin. Stick to Arial font, 16 points for section titles and 12 points for body. Make and justify all sound assumptions that you might require. The deliverable should contain the following structure:

1. Introduction - clearly answering all questions for items **A**, **B** and **C** depicted in Section 1.
2. Functional architecture - main components (i.e., modules) and interactions between them:
    - Description of the functionalities of each module and its purpose.
    - Functional architecture diagram (boxes and arrows).
3. Tool selection
    - For each of the components in your functional architecture discuss which of the studied tools (e.g., HDFS, HBase, MongoDB, Spark, ...) could implement it. If no existing tool fits your component, clearly discuss why.
        - You can follow the matrix approach to compare different tools (see provided reference). If you follow this approach, provide all matrices in an appendix.
4. Use cases and data flows (minimum two for data ingestion/dispatch/storage and two for data processing/analysis) - describe how data are processed throughout your architecture.
    - What components are involved? What is the input/output in each? What are the applied transformations?
    - They must be depicted using one of the following formalisms:
        - ETL processes
        - BPMN diagrams
        - Sequence diagrams
    - Performance and quality metrics
        - Performance metrics such as scalability, throughput, etc.
        - Quality metrics such as accuracy, completeness, timeliness, consistency, etc.
5. PoC description
    - Ideal infrastructure setting - present the ideal implementation of your architecture within a distributed cluster. What is the optimal size of the cluster? How would you distribute each of the components within the available machines?
    - Programming languages and external libraries being used.

In LearnSQL you will find some useful references that should help you devising the architecture and the tool selection.

5 Timeline

We will adhere to the following timeline, refer to the course's website for the specific dates.

1. Optional session to clarify doubts. You are highly encouraged to send in advanced (by email) your questions.

2. Online submission
3. Feedback session
4. Public presentation (including PoC demo)

After the feedback session, a mark will be generated which accounts for 40% of the weight. The remaining 60% will be generated after the presentation.

6 Evaluation criteria

1. Conciseness
   - Report – max 10 pages
   - Appendix 1 (optional): tool selection matrices (excluding any discussion) – max 2 pages
   - Appendix 2 (optional): data flow figures (excluding any discussion) – max 2 pages
2. Understandability
3. Coherence
4. Soundness