

## Project Components:

The topics for the DWH project for Fall 2020 semester at IBA have already been assigned to the students. See the file [here](#).

The project is divided into different tracks. A student needs to select one track to implement the project. First, let me define the components of the tracks.

**Component 1: Acquire Business Domain Knowledge (5%):** In this component, the student has to fill in the business knowledge template regarding the assigned project. The student can choose to leave out one or more tabs in this template, or add his/her own. In other words, demonstrate your understanding of the Business Domain in any way that you think is the best, e.g., maybe tweets etc. are not important for you so you can leave them out.

**Component 2: Generating the OLTP Database Diagram (12%):** Based on your business knowledge, figure out the important entities, their attributes and inter-relationships. Then, generate the database diagrams with proper cardinalities and normalization (up to 3rd Normal Form). You can first generate the ERD for facilitation and then the database diagram. Please generate all possible tables and all possible attributes within tables. You can make assumptions. Output Required: I need to see the database diagram, description of attributes, and relationships between tables.

*NB:* I am asking you to generate the database diagram because I was unable to find any complete diagram for my project ideas, or get it from the industry (for so many projects). The best thing is to let you have your own independence over the matter. It is also a bit strange to modify some downloaded database schema to suit your own needs. Overall, designing your own DB (for a DWH project) seems to be a reasonable idea.

**Component 3: Generating the Dimension Models (27%):** Generate one or more dimensional models by filling up *all the tabs* of the dimensional modeling Excel file. Bonus marks to be assigned for more than one data mart. If you have done Component 2 in detail, then this component will not be that difficult.

**Component 4: Data Lakes (5%):** What are the core components of a data lake? For the answer, compare the data lake offerings of AWS, Azure and IBM and also Apache Hudi along with any other open source solutions you can find (plus Wikipedia and YouTube). Output Required: Give as much detailed response as possible in your own words.

**Component 5: Acquiring and/or Generating Data (10%):** You need to generate, or acquire data related to your assigned project. There are many data repositories online,

particularly Kaggle, UCI Machine Learning Repository, <https://www.dataquest.io/blog/free-datasets-for-projects/>, <https://catalog.data.gov/dataset>, <https://github.com/awesomedata/awesome-public-datasets> and <https://www.mygreatlearning.com/blog/free-download-datasets/>. You can generate synthetic data from Python code as well. The data volume should definitely be in Gigabytes. Output Required: Details of the data, and how it was acquired/generated (also the data generation code if used).

**Component 6: Dimensional Modeling with Hive (23%):** Generate tables over Hive to facilitate your dimensional modeling requirements, import the generated/acquired data into these tables, and then execute them to show the results. Output Required: I need to see all the executed commands with snapshots of the more important ones.

**Component 7: Dimensional Modeling with Spark (23%):** Generate data frames over Spark to facilitate your dimensional modeling requirements, import the generated/acquired data into these tables, and then execute them to show the results using Spark SQL. Output Required: I need to see all the executed commands with snapshots of the more important ones.

### Project Tracks:

- Track 1
  - Component 1
  - Component 2
  - Component 3
  - Component 4
- Track 2:
  - Component 1
  - Component 2
  - Component 5
  - Component 6
- Track 3:
  - Component 1
  - Component 2
  - Component 5
  - Component 7

Progress Report Submission: 28<sup>th</sup> December, 2020

Project Final Report Submission: 10<sup>th</sup> January, 2020

#### Submission Requirement:

- Business Knowledge Template (For all Tracks)
- Dimensional Modeling Workbook (for Track 1)
- Word and PDF file containing a short report of the whole project (for Track 1), e.g., details of acquiring the business knowledge, comments on this knowledge, assumptions in designing the ERD or Database diagram, any details or assumptions of dimensional modeling etc.
- Word and PDF file detailing the data generating / data acquiring process (for Track 2 and Track 3).
- Generated/Acquired data (in CSV format) (for Track 2 and Track 3)
- Word and PDF file showing the series of all Hive commands executed for dimensional modeling along with snapshots of outputs (For Track 2)
- Word and PDF file showing the series of all Spark SQL commands executed for dimensional modeling along with snapshots of outputs (For Track 3)

#### Important:

- Students attempting Track 2 and Track 3 will get a small bonus over candidates attempting Track 1