POLITECNICO
MILANO 1863

Project - Delivery III

Andrea Tocchetti
andrea.tocchetti@polimi.it

Project Delivery

Details about the second project delivery

- You have **“about” two weeks** to deliver the second part of the project, starting **tomorrow**
  - Spark
- Continue to expand the document with the 1st and 2nd delivery
- You will be asked to deliver the 3rd delivery on WeBeep by the **15th of December** (actual deadline 16th of December at **1:00 A.M.**).
- **“about”** in a sense that the material should be ready to be presented at the final presentation that will be held on the 13th of December but you can deliver it at a later time (15th of December).

Project Delivery

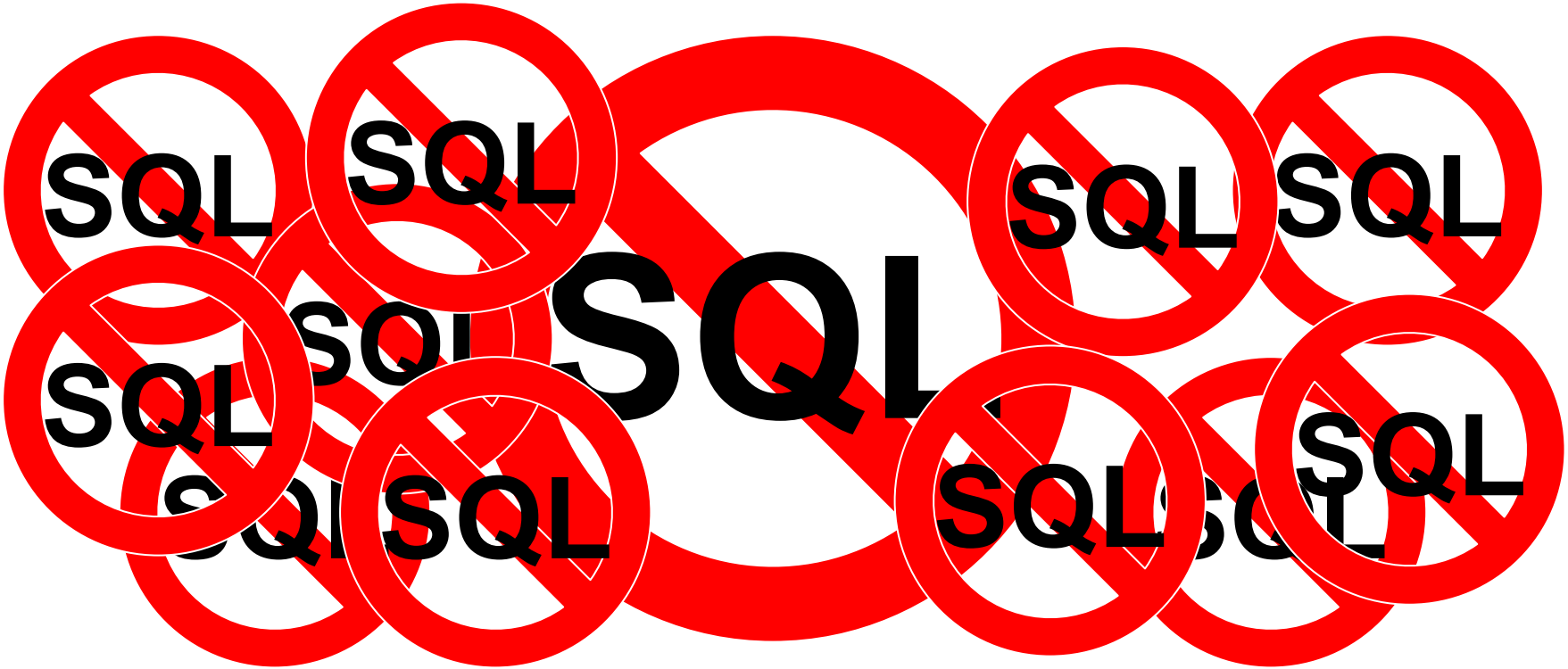
The final delivery should contain

- A description of the structure of your dataset in pyspark
  - Use at least 2k - 3k tuples
- The script used to transform the data in the pyspark format **WITH COMMENTS**, if needed
- The final script to perform all the operations **WITH COMMENTS** on every query/important operation performed.
- A screenshot of the query or its text with its explanation (for each query)
- A screenshot of the result of the query (for each query)

Queries

- Perform at least 5 data creation/update operations (add a new row, remove a row, etc.)
- Perform at least 10 queries with the following complexities (**they are provided using their SQL equivalents for simplicity**)
  - WHERE, JOIN
  - WHERE, LIMIT, LIKE
  - WHERE, IN, Nested Query
  - GROUP BY, 1 JOIN, AS
  - WHERE, GROUP BY
  - GROUP BY, HAVING, AS
  - WHERE, GROUP BY, HAVING, AS
  - WHERE, Nested Query (i.e., 2-step Queries), GROUP BY
  - WHERE, GROUP BY, HAVING, 1 JOIN
  - WHERE, GROUP BY, HAVING, 2 JOINs

Queries




ANY
Questions?