Lab Assignment 3

Problem:

How to leverage confidential data from different stakeholders to meet shared sustainability targets (Creation of a data lake)?

Key Contributions Criteria

The aim is to develop a data sharing systems or Database building blocks that:

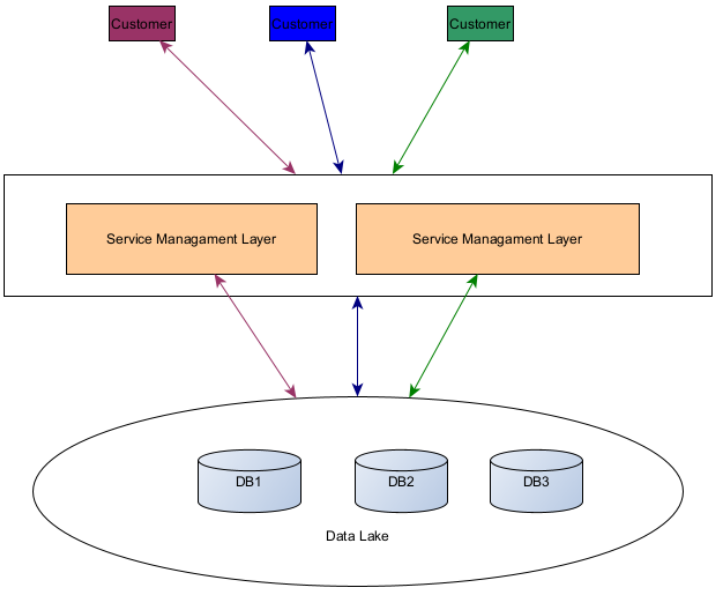
- Provide secure federated data and information sharing
- Ensure that the integrity of operational data from each organization / site is not compromised
- Does not reveal confidential company information or allow organizations / individuals to be identified
- Allow the data / information to be processed and analysed by common tools
- Ensure that the confidentiality of data shared by each stakeholder is maintained: e.g. data shared cannot be back processed for reverse engineering, etc…
- Create trust and collaboration between organizations by ensuring shared data cannot be used for anything other than the common goals, and this is transparent and verifiable by the data contributors.

Possible Approaches

The proposed solution might include, but not limited to:

- Algorithms for encryption, distributed learning, prediction task mechanism, coupled with methodologies for ensuring the data integrity from all sources.
- Federated database system (FDBS), Open Data platforms, blockchain, distributed optimization, decomposition methods, game theory, market design, cryptography, coupled with methodologies for ensuring the data integrity from all sources.
- Novel approaches to prove data integrity and the approved use of the data to the data contributors.
- You can think of exploring available cloud services from different service providers.

Possible Architecture



The data lake comprises of multiple databases (more than one). No of customers are more than one.

The service management layer takes care of data access of each customer. Service Management layer will identify the customer, its data, and stores in appropriate database (DB1 or DB2 or DB3) based on the type of data. Similarly, when the customer reads the data, it identifies the location of the data of that customer and sends it back to that customer. Service management layer basically should ensure data transparency. The different coloured arrows basically mean this.

You are free to make further assumptions as required.

Deliverables

The deliverables include:

- Detailed project Report (Framework designed, Technology stack used, Datasets used, Snapshots of the implementation).
- Also mention specific observations such as viability, challenges, etc.
- Presentation of the project.

Rubric

The evaluation includes (10%):

- Data Lake implementation (10 marks)
- Sharing mechanism and confidentiality (15 marks).
- Presentation and Report (5 marks).