**Holistic Lab Execution Environment Data Processing With EDL 2.0**

The project **Holistic Lab Execution Environment data processing with EDL 2.0** focuses on experimental data of a pharmaceutical company and perform transformations on data. Holistic lab execution environment data hold information regarding various experiments, chemical compositions of drugs, batch id, status etc. The amount of data generated in this lab will be huge and will be real streaming. Enterprise data lake Operations will source data from HLEE ELN system via REST services provided by Biovia. EDL will continuously ingest HLEE data from HLEE ELN system with a time gap of about 15 minutes from the time the experiment message is available for ingestion by EDL, where the ingestion and processing time will be dependent on the size of the file and the number of concurrent files ingested at a particular time. The EDL Operations System shall maintain process logs which includes error messages and trace back for all steps executed as part of HLEE near real time data ingestion process. A type of HLEE data is HLEE C&C data it hold experiment recipes.

Enterprise Data Lake (EDL) is a large repository for many different types and sources of data, they can be structured or unstructured, internal or external, to facilitate different ways of accessing and analyzing the data. It stores relational data from the line of business applications, and non-relational data from mobile apps, IoT devices, and social media. Data Lake ingests data from many systems such as CRM, ERP and other transactional systems. This rich data ecosystem can now support combining multiple sources of data for more accurate analytics and never-before-possible insights into business operations.

The project HLEE data processing with EDL 2.0 stores data in the form of json format, and for the further analysis, each data will be extracted and finally it will be stored in Postgres database and redshift database. To provide reports on visualization tools including custom made user interfaces, Reporting tools like TIBCO Spotfire and Tableau. The security of data is ensured by using different technologies such as Kerberos authentication, IAM (Identity Access Management). Different key resources used by the system are GIT repository, Airflow, Databricks, Postgres Database, Amazon S3, ISM/ServiceNow (for service request/incident management).

Akza John

Roll no-4