ABBVIE CURATION PROCESS FORM

TABLE OF CONTENTS

1. PROJECT OVERVIEW

Name of the Client: AbbVie

Name of the Project: Curation

Overview of the Project:

The AbbVie Curation team offers better organization and visibility to the critical information useful for the next innovations in academic, clinical, pre-clinical, and industrial research by providing hybrid networks (SQL, AWS, Azure). The team is also responsible for applying normalization, knowledge extraction and loading the data into neo4j then deploying the changes into DEV, int, QA and prod environments.

This process is designed to help researchers prepare a research project for archiving in accordance with FAIR principles, ensuring the projects value is preserved during and after project completion. Data curation may seem like a daunting task, but this process aims to simplify the process by reducing it into six themed steps, **CURATE:**

- **C**heck Files (Unstructured (Files) and Structured (Tables)).
- **U**nderstand the data and external constraints.
- **R**equest (or locate) missing information.
- **A**ugment metadata for findability.
- **T**ransform file formats for reuse.
- **E**valuate for FAIRness.

2. BUSINESS OBJECTIVE

Enhance the enterprise data transformation through the enactment of governance norms and regulations.

3. REQUIREMENTS

3.1 Skill Sets Required

- Apache Spark
- Knowledge on Almaren framework
- SQL
- Knowledge on BOTS framework
- Knowledge on Nabu platform
- PySpark
- Neo4j
- AWS
- Java
- Python

3.2 Tools/Technologies Required

- PgAdmin/DBeaver
- Nabu UI
- Cloudera
- MobaXterm
- Git
- Pipeline Pilot Professional Client
- Visual Studio code
- Visual Studio Enterprise
- IntelliJ
- Zeplin

4. REFERENCES

4.1 Documents for Reference

- Almaren framework
- Apache Spark
- Cloudera
- Neo4j

4.2 Links for Reference

Sample CASE statements: https://www.postgresql.org/docs/7.4/functions-conditional.html

Almaren framework: https://github.com/modakanalytics/almaren-framework

Apache Spark: https://spark.apache.org/docs/latest/

Reading Spark tables using sourceSql: https://github.com/modakanalytics/almaren-framework#sourcesql

Neo4j connector for establishing relationships between entities: https://github.com/music-of-the-ainur/neo4j.almaren

Basic introduction on Neo4j: https://neo4j.com/docs/getting-started/current/

Graph database concepts in Neo4j: https://neo4j.com/docs/getting-started/current/graphdb-concepts/

Basic introduction to PySpark:https://spark.apache.org/docs/latest/api/python/index.html

5.Training to be completed and software's to be installed after getting VM access

- Complete the iso trainings pending by logging into the isotrain website using the credentials.
- Navigate to the AbbVie self-service portal and click on the tray icon where it redirects to the page where the ticket has to be raised to install the software's required.
- Raise the ticket for the following software's which are required to installed
  - PgAdmin 4 v4
  - BIOVIA Pipeline Pilot Client
  - MobaXterm v11.0

Note:

- Complete the Iso trainings without fail with in the deadline
- Keep track on the AbbVie mail once requesting accesses are done

6. TASKS OVERVIEW

6.1 Description of tasks to be performed

The team is responsible for transforming the raw data into a standardized format as per the industrial standards using SQL, store it in a data frame, write that data frame into an s3 bucket by deploying a pipeline in Nabu. Extra columns such as strength and confidence which are used to measure the efficiency of the drug can be added to the standardized tables as per the client's requirement. The team is responsible for creating entities and publish the data by building relationships between different entities using Neo4j, a graph data platform.

6.2 Tasks checklist

- Normalize the source data according to the industrial standards by writing case statements in SQL, execute using almaren framework and store it in a data frame.
- Writing the data frame into an S3 bucket in a parquet format
- Deploy the pipeline in Nabu platform
- Add the extra columns such as strength and confidence which are used to measure the effectiveness of the drug out of the standardized data
- Establishing relationship in Neo4j using the Neo4j connector in Almaren framework.