ANALYTIXLABS

Data Engineering – Capstone Project-1

Data Engineering Capstone Project-1

Introduction:

- You have been hired as a new data engineer at Analytixlabs. Your first major task is to work on data engineering project for one of the big corporation's employees data from the 1980s and 1995s. All the database of employees from that period are provided six CSV files. In this project, you will design data model with all the tables to hold data, import the CSVs into a SQL database, transfer SQL database to HDFS/Hive, and perform analysis using Hive/Impala/Spark/SparkML using the data and create data and ML pipelines.

Project Description:

In this project, you are required to create end to end data pipeline and analyzing the data.

Technology Stack:

you are required to work on below technology Stack.

- MySQL (to create database)
- Linux Commands
- Sqoop (Transfer data from MySQL Server to HDFS/Hive)
- HDFS (to store the data)
- Hive (to create database)
- Impala (to perform the EDA)
- SparkSQL (to perform the EDA)
- SparkML (to perform model building)

Data Engineering Capstone Project-1

Project Objective: As part of this project, you are required to work on

1. Create data model as per your understanding from the data (you are required include tables names, relation between tables, column names, data types, primary & foreign keys etc.)
Tip: You can create ER diagram either in EXCEL or using the link https://www.quickdatabasediagrams.com/ (Preferably in this app)

2. Create database & tables in MySQL server as per the above ER Diagram

3. Create Sqoop job to transfer the data from MySQL to HDFS (Data required to store in Parque/Avro/Json format)

4. Create database in Hive as per the above ER Diagram and load the data into Hive tables

5. Work on Exploratory data analysis as per the analysis requirement using Impala and Spark SQL (expecting to get the data from hive tables).

6. Build ML Model as per the requirement.

7. Create entire data pipeline and ML pipe line

8. Upload the entire project repository including source code to Github (you are required to create github account if you don't have it)

Data Description

Data Description: Please find the details of all the tables

a. Titles (titles.csv):

title_id – Unique id of type of employee (designation id) – Character – Not Null
title – Designation – Character – Not Null

b. Employees (employees.csv):

emp_no – Employee Id – Integer – Not Null
emp_titles_id – designation id – Not Null
birth_date – Date of Birth – Date Time – Not Null
first_name – First Name – Character – Not Null
last_name – Last Name – Character – Not Null
sex – Gender – Character – Not Null
hire_date – Employee Hire date –Date Time -Not Null
no_of_projects – Number of projects worked on – Integer – Not Null
Last_performance_rating – Last year performance rating – Character – Not Null
left – Employee left the organization – Boolean – Not Null
Last_date - Last date of employment (Exit Date) – Date Time

Data Description

Data Description: Please find the details of all the tables

c. Salaries (salaries.csv):

emp_no – Employee id – Integer – Not Null
Salary – Employee's Salary – Integer – Not Null

d. Departments (departments.csv)

dept_no - Unique id for each department – character – Not Null
dept_name – Department Name – Character – Not Null

e. Department Managers (dept_manager.csv)

dept_no - Unique id for each department – character – Not Null
emp_no – Employee number (head of the department) – Integer – Not Null

f. Department Employees (dept_emp.csv)

emp_no – Employee id – Integer – Not Null
dept_no - Unique id for each department – character – Not Null

Exploratory Data Analysis

The queries in database include

1. A list showing employee number, last name, first name, sex, and salary for each employee
2. A list showing first name, last name, and hire date for employees who were hired in 1986.
3. A list showing the manager of each department with the following information: department number, department name, the manager's employee number, last name, first name.
4. A list showing the department of each employee with the following information: employee number, last name, first name, and department name.
5. A list showing first name, last name, and sex for employees whose first name is "Hercules" and last names begin with "B.“
6. A list showing all employees in the Sales department, including their employee number, last name, first name, and department name.
7. A list showing all employees in the Sales and Development departments, including their employee number, last name, first name, and department name.
8. A list showing the frequency count of employee last names, in descending order. (i.e., how many employees share each last name
9. Histogram to show the salary distribution among the employees
10. Bar graph to show the Average salary per title (designation)
11. Calculate employee tenure & show the tenure distribution among the employees
12. Perform your own Analysis (based on the data understanding) – At least 5 additional analysis

Spark ML Model

You need to build binary classification model in spark ML

a. Read the data from Hive tables
b. You required to join all the tables at employee level
c. Build classification model using few algorithms (like logistic regression or random forest etc…) in spark ML (By considering Target variable is left, other variables are independent variables)
d. Create ML pipeline

Expectations from Trainees

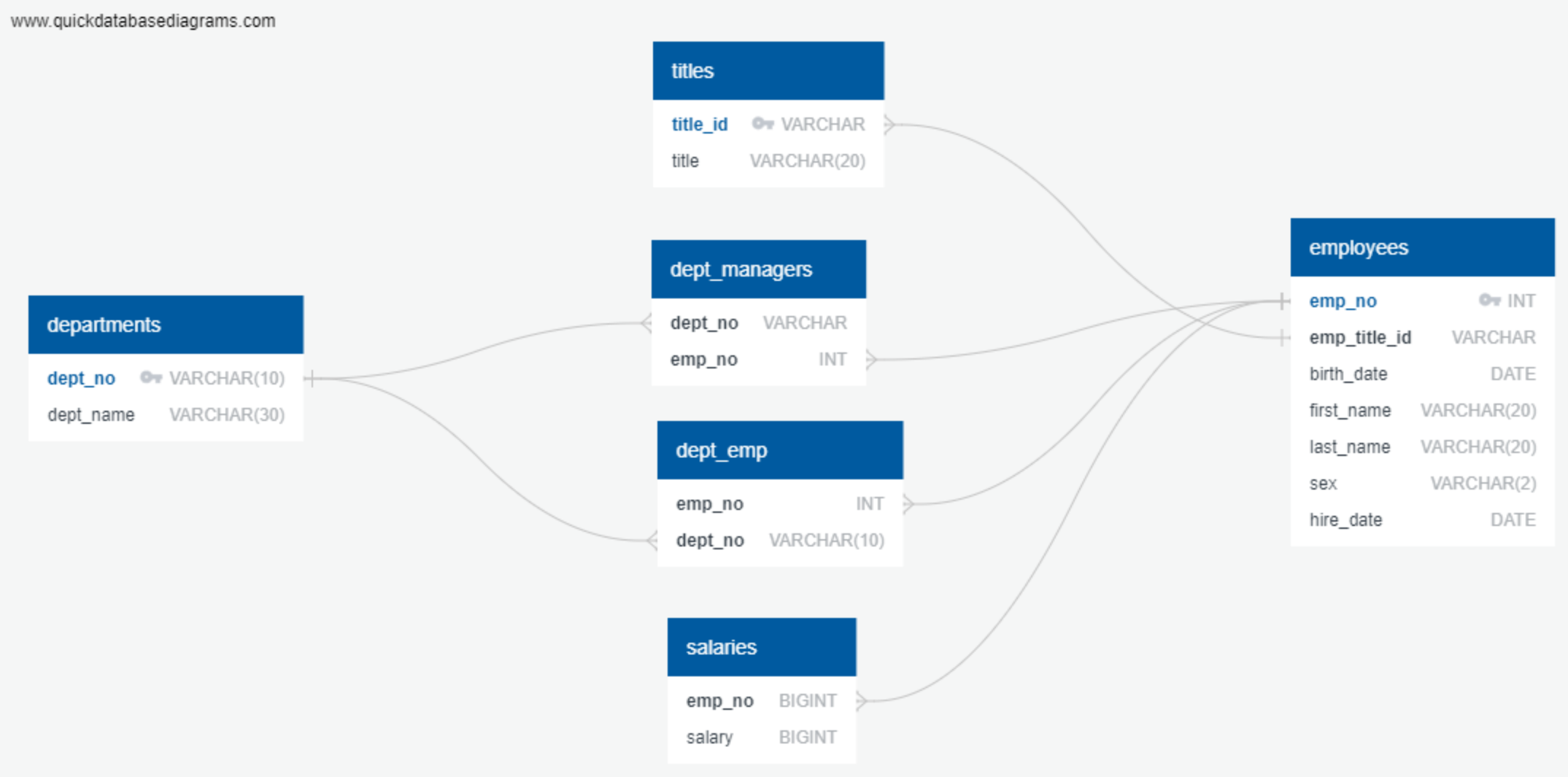
You are required to work and submit the Github Repository link with below details.

a. Image file of your ERD
b. Submit all code files .sh/.sql/.hql/.ipynb with proper comments.
c. You are required to work on end to end pipeline with proper documentation with below details (word/Excel/Ppt)
    a. Business objective
    b. Data used & description
    c. Technology stack used
    d. ER Diagram (data model)
    e. Architecture of pipeline (stages)
    f. Outputs for different analysis (EDA & ML Model)
    g. Challenges you faced
    h. Next steps
d. Create and upload a repository with the above files to GitHub and share the link with Chandra.
e. Ensure your repository has regular commits and having README.md file.

Time Lines: On or before End of Thursday (19th May 2022)

Appendix

Sample ERD for your reference

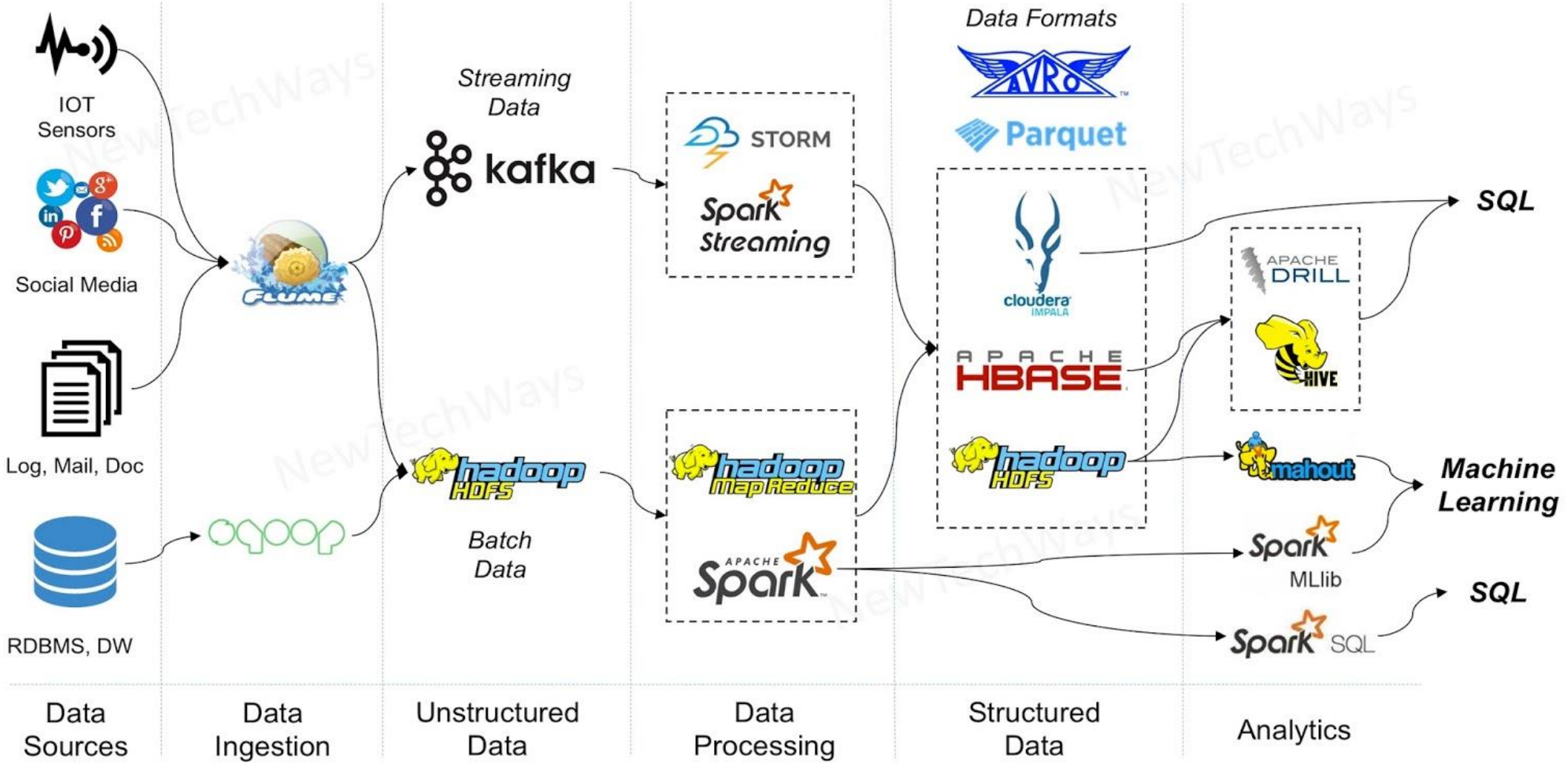


Sample code for table creation

```
CREATE TABLE employees (
    emp_no INT NOT NULL,
    emp_title_id VARCHAR NOT NULL,
    birth_date DATE NOT NULL,
    first_name VARCHAR(20) NOT NULL,
    last_name VARCHAR(20) NOT NULL,
    sex VARCHAR(2) NOT NULL,
    hire_date DATE NOT NULL,
    PRIMARY KEY (emp_no),
    FOREIGN KEY (emp_title_id) REFERENCES titles(title_id)
);

select * from employees
```

Sample architecture of Pipeline



Contact us

Visit us on: http://www.analytixlabs.in/

For course registration, please visit: http://www.analytixlabs.co.in/course-registration/

Email: info@analytixlabs.co.in

(or) Call us we would love to speak with you: (+91) 88021-73069

Join us on:

Twitter - http://twitter.com/#!/AnalytixLabs

Facebook - http://www.facebook.com/analytixlabs

LinkedIn - http://www.linkedin.com/in/analytixlabs

Blog - http://www.analytixlabs.co.in/category/blog/

Gurgaon Address:
GF 382, Sector 29,
Adjoining IFFCO Metro Station (Gate 2),
Next to Vasan Eye Care Hospital,
Gurgaon, Haryana 122001

Bengaluru Address:
Bldg 41, First floor,
14th Main Road, Sector 7, HSR Layout
Bengaluru - 560102
Landmark: Max store

Noida Address:
First Floor, A-78,
A Block, Sector 2, Noida,
Uttar Pradesh - 201301
(Adjacent to Sector 15 metro station)