Data Engineering
Capstone Project-1

Business Objectives

- Exploratory Data Analysis of provided Data Set.
- Using the data set to come up with meaningful insights.
- Build Predictive Model

Technology Stack Used

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)

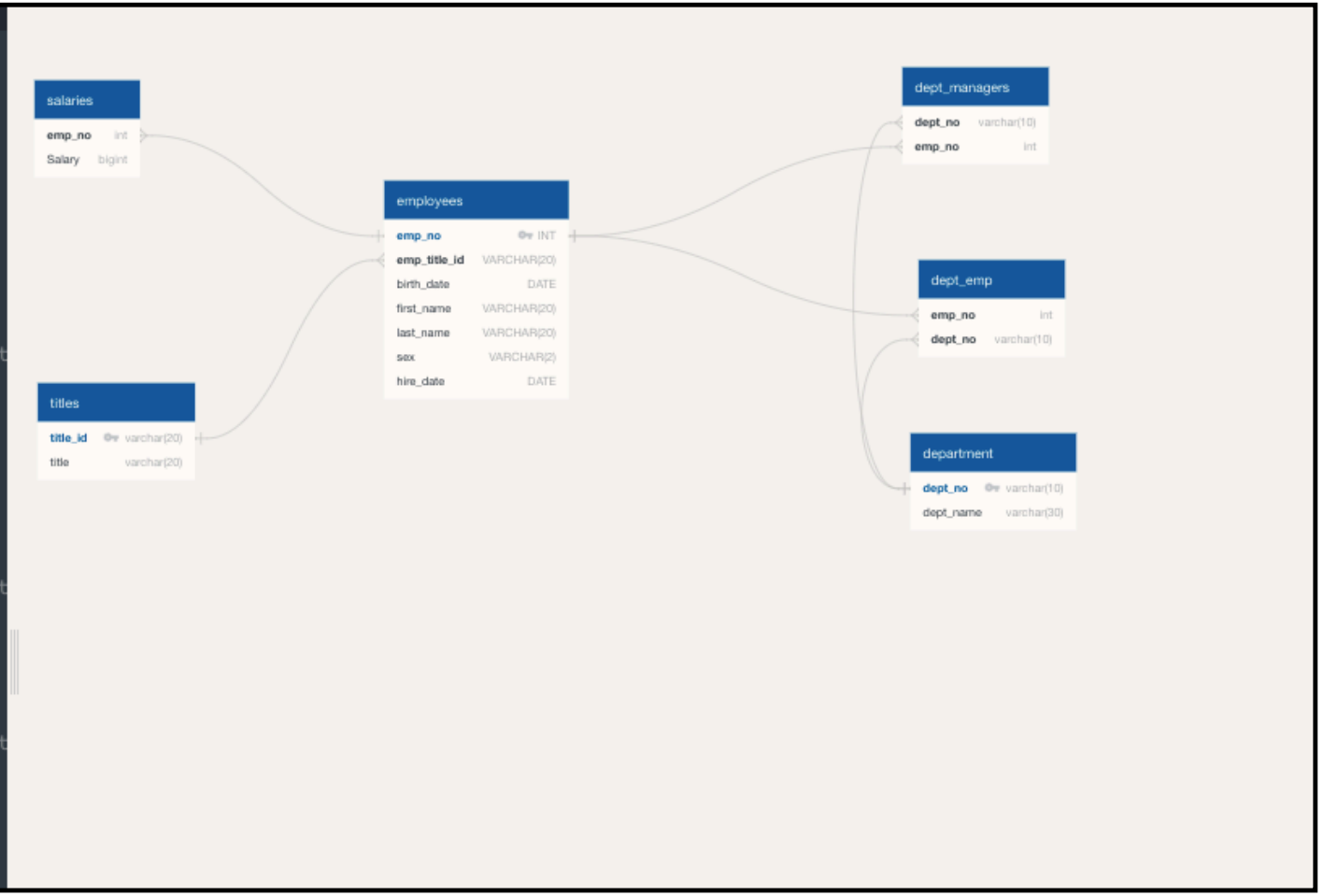
Data Set Used

Various CSV used:

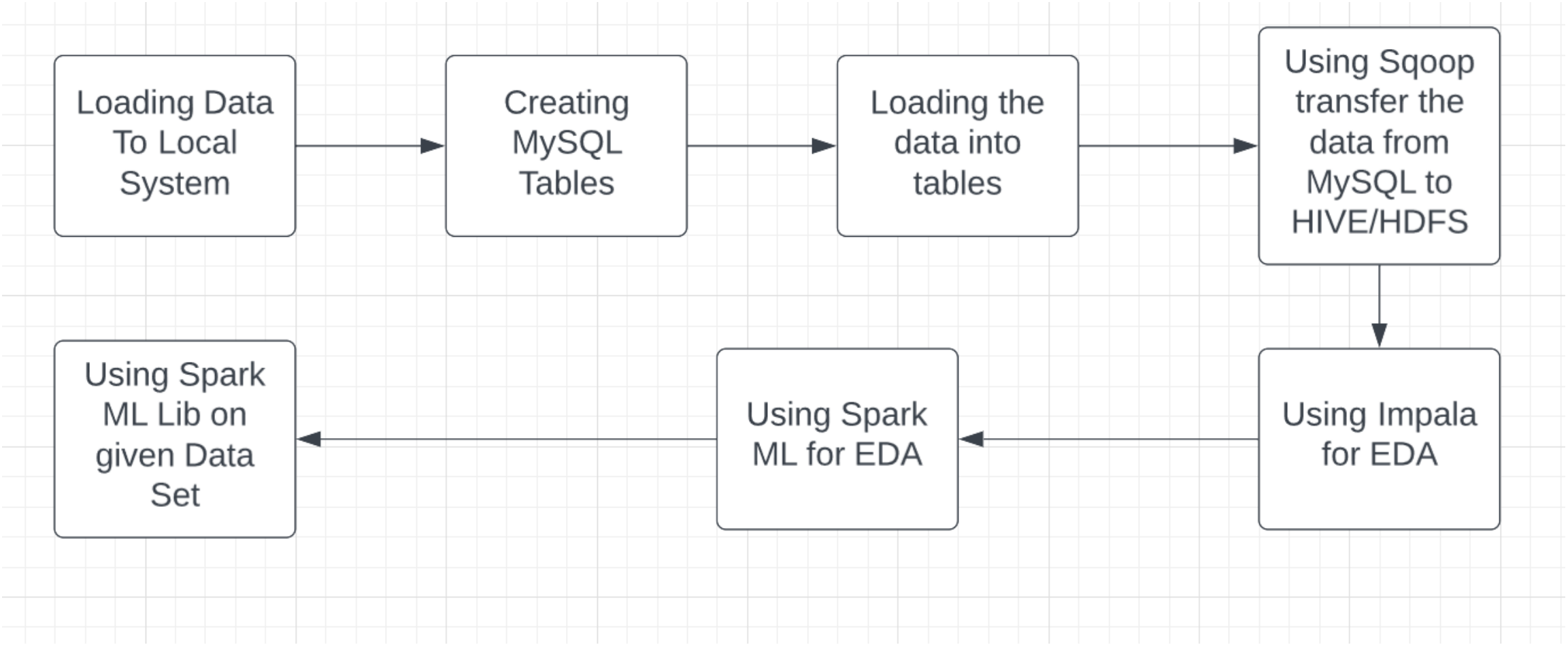
- Departments
- Department Managers
- Department Employees
- Employees
- Salaries
- Titles

ERD

```
department
--
dept_no varchar(10) PK
dept_name varchar(30)

titles
--
title_id varchar(20) PK
title varchar(20)

employees
--
emp_no INT PK
emp_title_id VARCHAR(20) FK >- titles.tit
birth_date DATE
first_name VARCHAR(20)
last_name VARCHAR(20)
sex VARCHAR(2)
hire_date DATE

dept_managers
--
dept_no varchar(10) FK >- department.dept
emp_no int FK >- employees.emp_no

dept_emp
--
emp_no int FK >- employees.emp_no
dept_no varchar(10) FK >- department.dept

salaries
--
emp_no int FK >- employees.emp_no
Salary bigint
```

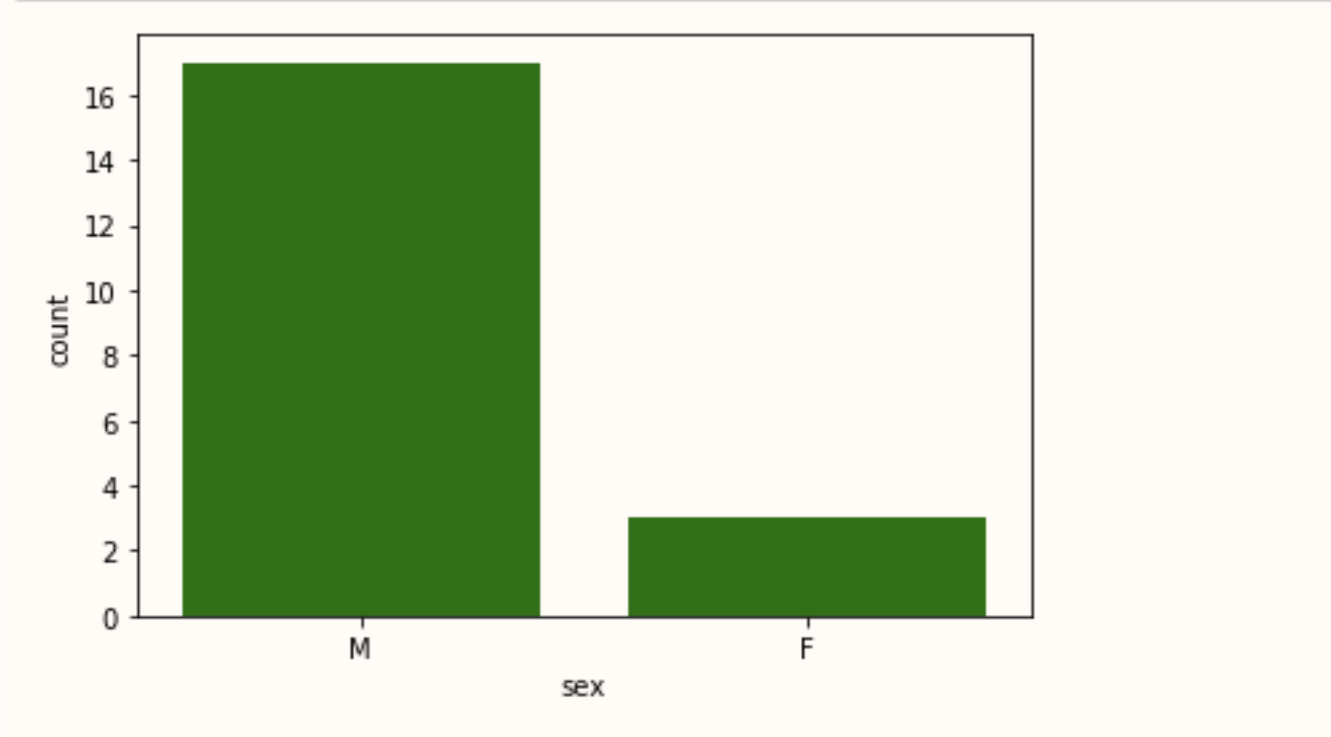


Pipeline Architecture

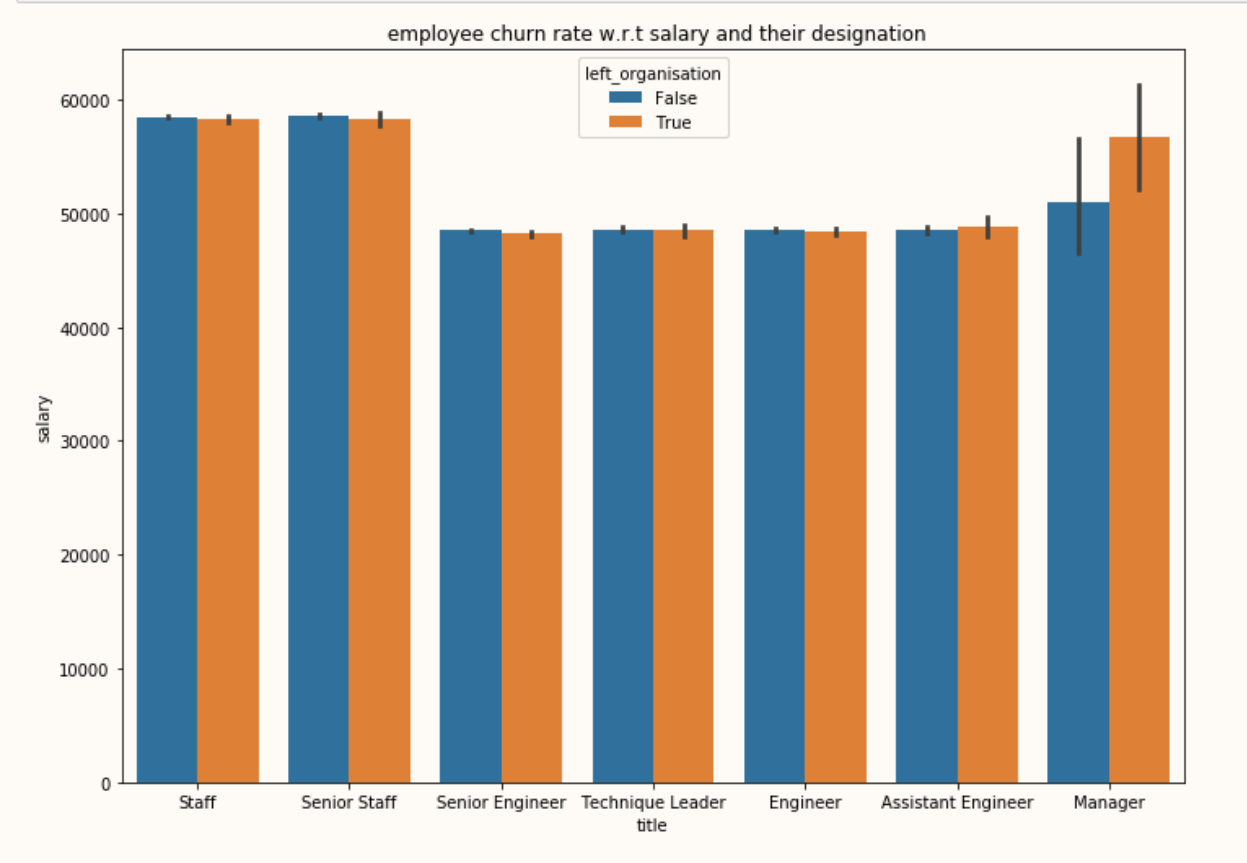


Outputs

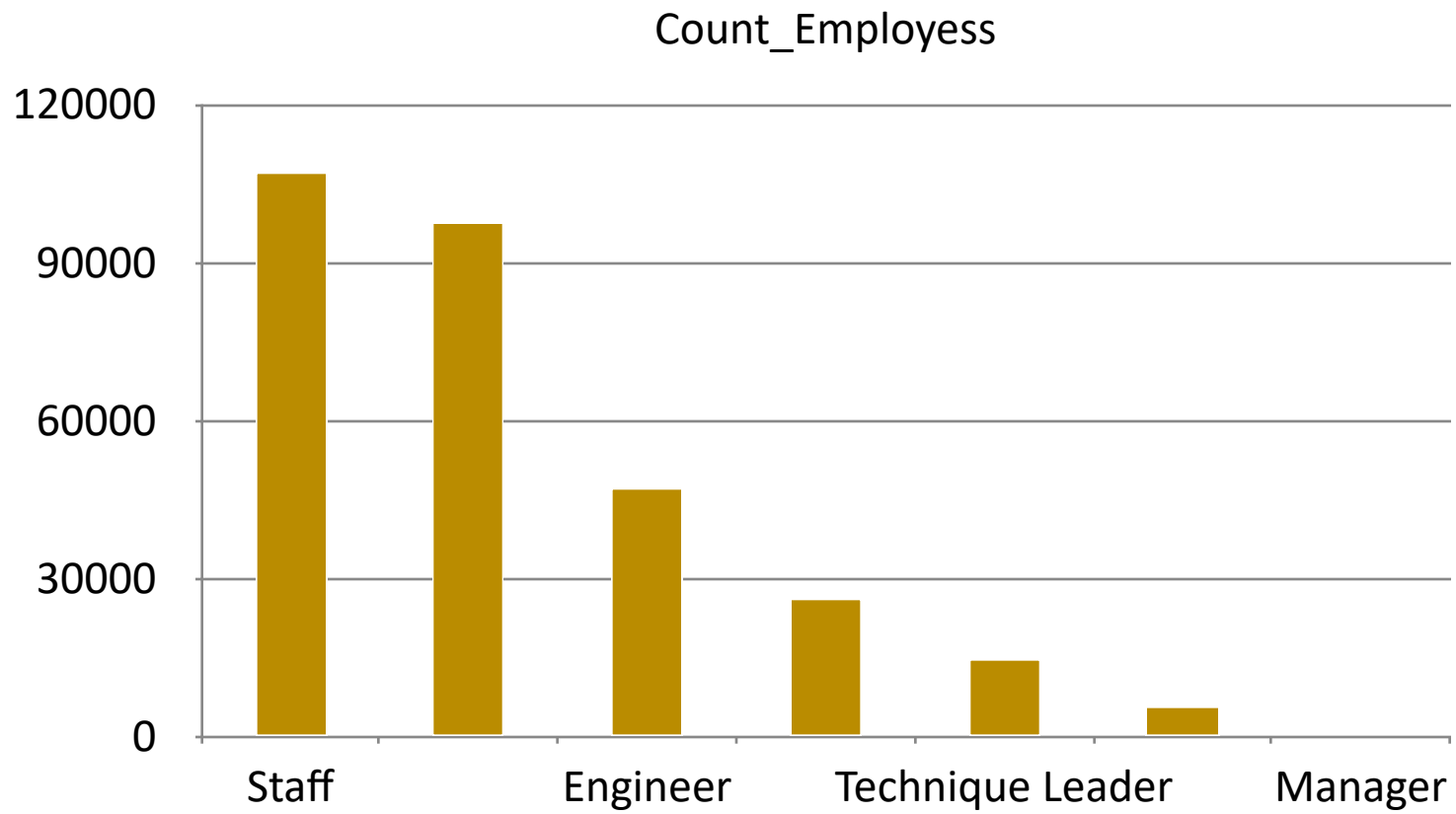
- top 20 Highest earning employees are shown on above table ; Gridswold Charmane is being one who is earning highest among all at 129492, Majority of top 20 highest earning employee are Male , there are only



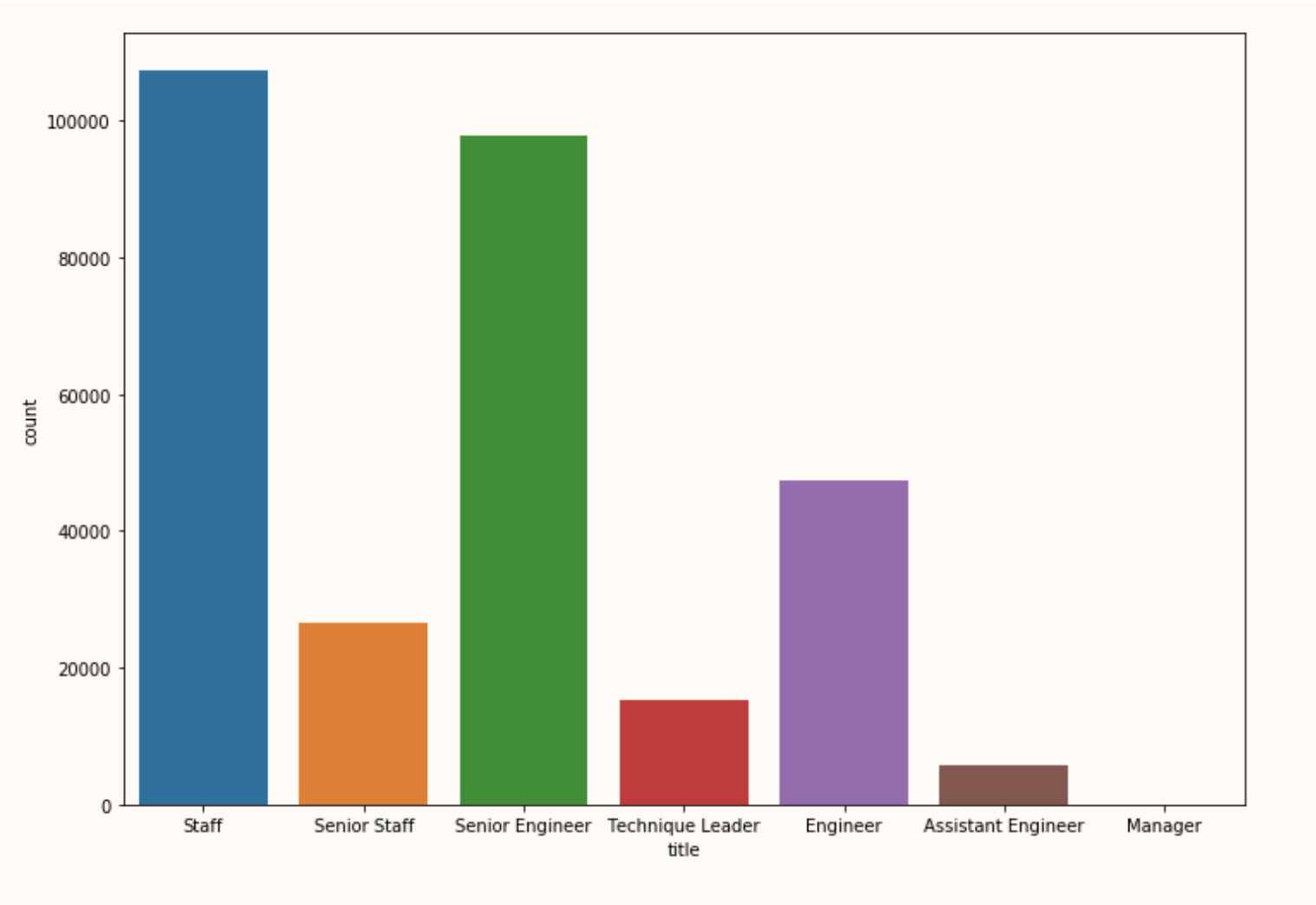
- According to given illustration , Majority of employees are in IT sector. Managers might churn regardingless of salary, There is high churn rate in manager designation at high salary.



- Title Distribution among the Employees.



- Title distribution



Challenges Faced

- Data Transfer using SQOOP.
- Building ML model
- Server

Steps Ahead

- Employee_Data_Analysis_exl

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

- Objective :
- There are three major steps for completing employee data analysis pipeline, each one has script file which uploaded here.

- ** STEPS **
- 1. Data Modeling , ER diagram
- - create mysql table
-

Steps Ahead

- - load csv file in created table
- - prepare script file createtb_sql.sql
- - source createtb_sql.sql
- 2.Data Engineering
- - create new directory in hdfs
- - import data from mysql to hdfs using sqoop
- - prepare script file shell_script.sh
- -sh shell_script.sh
-
- - create table in hive
- - load data into created table
- -prepare script hivefile
-
- 3. Analysis the data in impala
- 4. Import data in pyspark
- 5. EDA

Thank You