



Capstone Project 1

BIGDATA ENGINEERING

by
Subham Shit

Business Objectives

- Basic Exploratory Analysis on different metrics available to get insights at the Employee level as well as Department levels.
- To analysis the various factors connected to the Employee's Leaving and improve on those parameters viz Salary, Last Performance Ratings, Years of Service (Tenure), No of projects etc.
- Technology-wise Objective : **Creating an End-to-End Pipeline** to enhance the automation to do quick analysis and manage all the repetitive actions in a proper-structured manner.



Data Description

This dataset has total 6 tables (Records) –

1. Employees (300024),
2. Salaries (300024),
3. Titles (7),
4. Dept_Emp (331603 -> 300024),
5. Dept_Manager (24) and
6. Departments (9).

Data Description

1. Titles (titles.csv):

- a. **title_id** – Unique id of type of employee (designation id) – Character – Not Null
- b. **title** – Designation – Character – Not Null

2. Employees (employees.csv):

- a. **emp_no** – Employee Id – Integer – Not Null
- b. **emp_titles_id** – designation id – Not Null
- c. **birth_date** – Date of Birth – Date Time – Not Null
- d. **first_name** – First Name – Character – Not Null
- e. **last_name** – Last Name – Character – Not Null
- f. **sex** – Gender – Character – Not Null
- g. **hire_date** – Employee Hire date –Date Time –Not Null
- h. **no_of_projects** – Number of projects worked on – Integer – Not Null
- i. **Last_performance_rating** – Last year performance rating – Character – Not Null
- j. **left** – Employee left the organization – Boolean – Not Null
- k. **Last_date** – Last date of employment (Exit Date) – Date Time

3. Salaries (salaries.csv):

- a. **emp_no** – Employee id – Integer – Not Null
- b. **Salary** – Employee's Salary – Integer – Not Null

4. Departments (departments.csv)

- a. **dept_no** - Unique id for each department – character – Not Null
- b. **dept_name** – Department Name – Character – Not Null

5. Department Managers (dept_manager.csv)

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

6. Department Employees (dept_emp.csv)

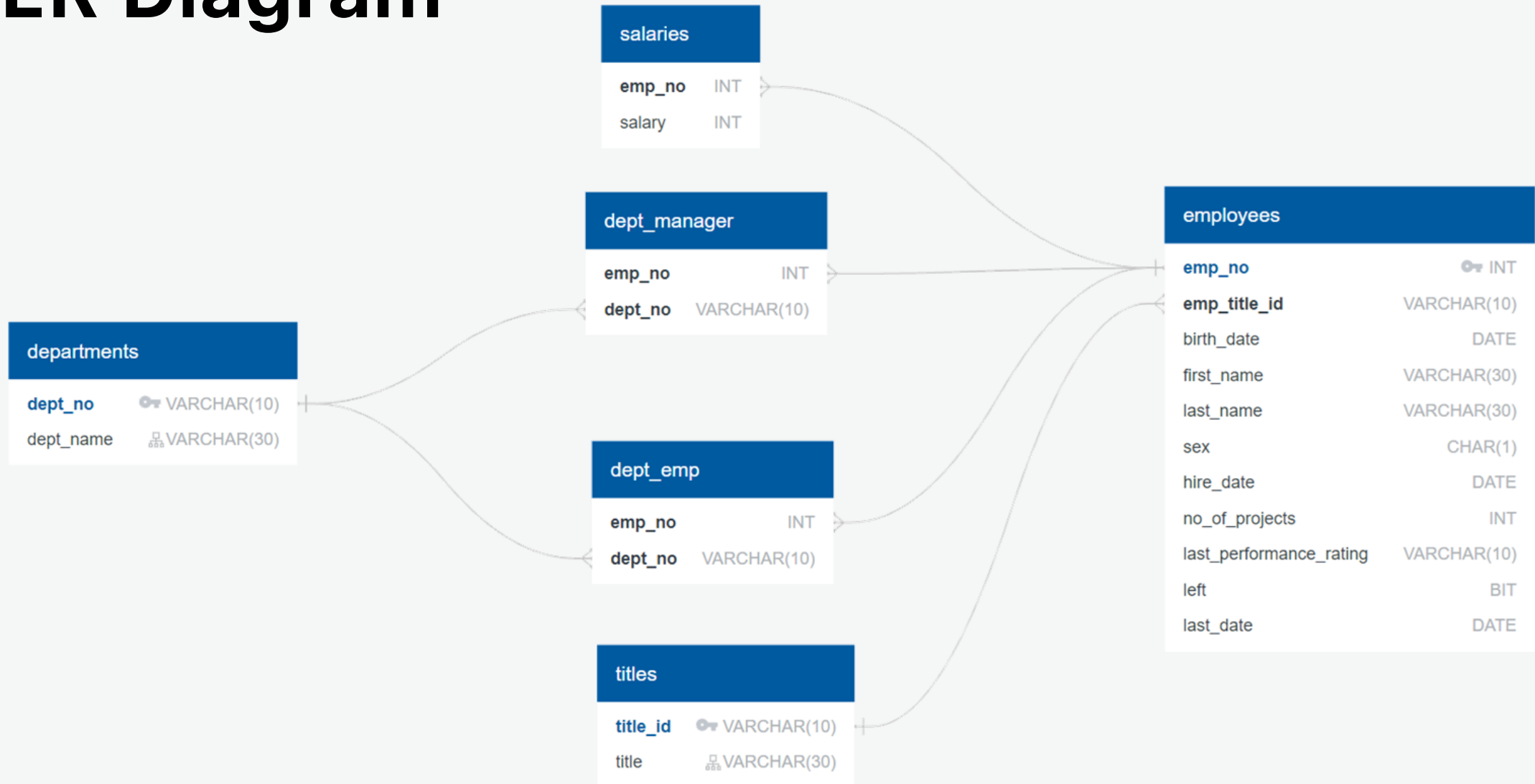
- a. **emp_no** – Employee id – Integer – Not Null
- b. **dept_no** - Unique id for each department – character – Not Null

Technology Stack

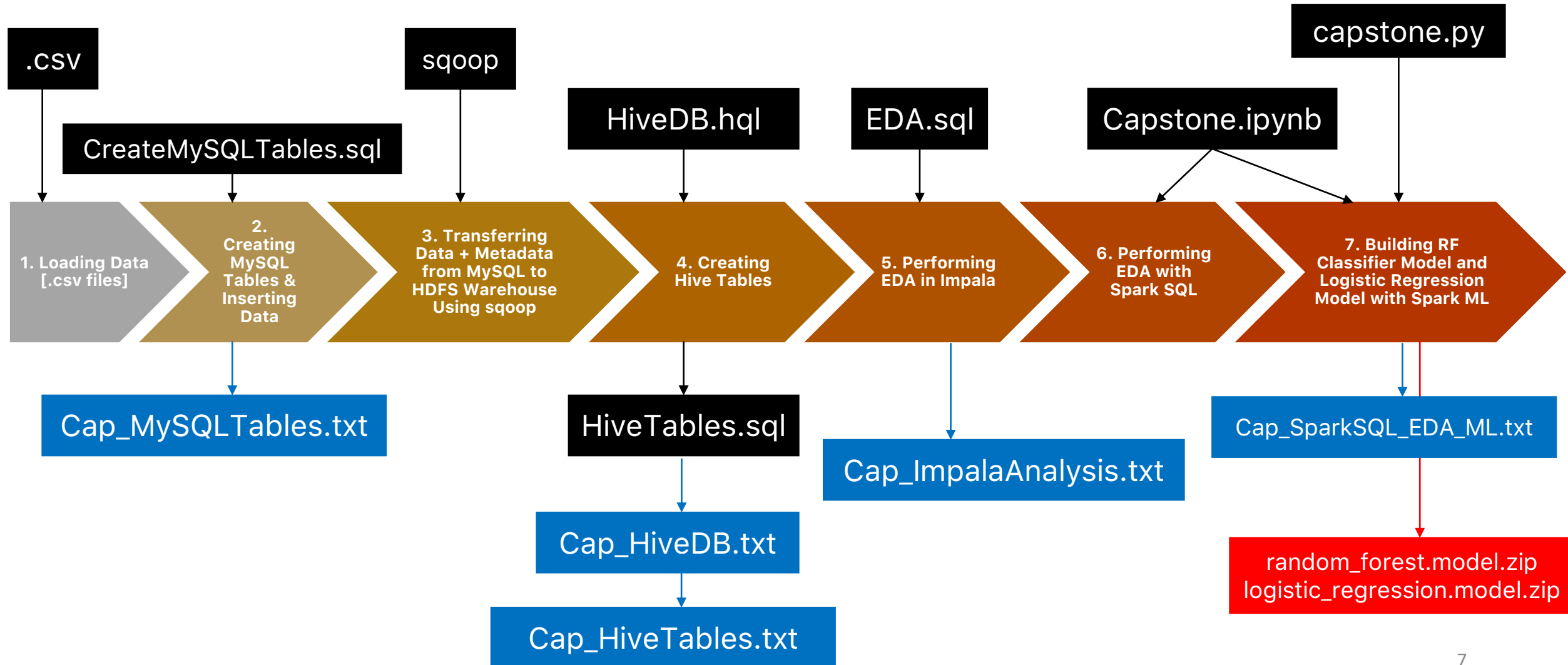
- **Linux Commands** (to run the bash script)
- **MySQL** (to create database - RDBMS)
- **Sqoop** (Transfer data from MySQL Server to HDFS/Hive)
- **HDFS** (to store the data)
- **Hive** (to create database)
- **Impala** (to perform the EDA)
- **SparkSQL & PySpark** (to perform the EDA)
- **SparkML** (to perform model building)
- **git** (for version control)



ER Diagram

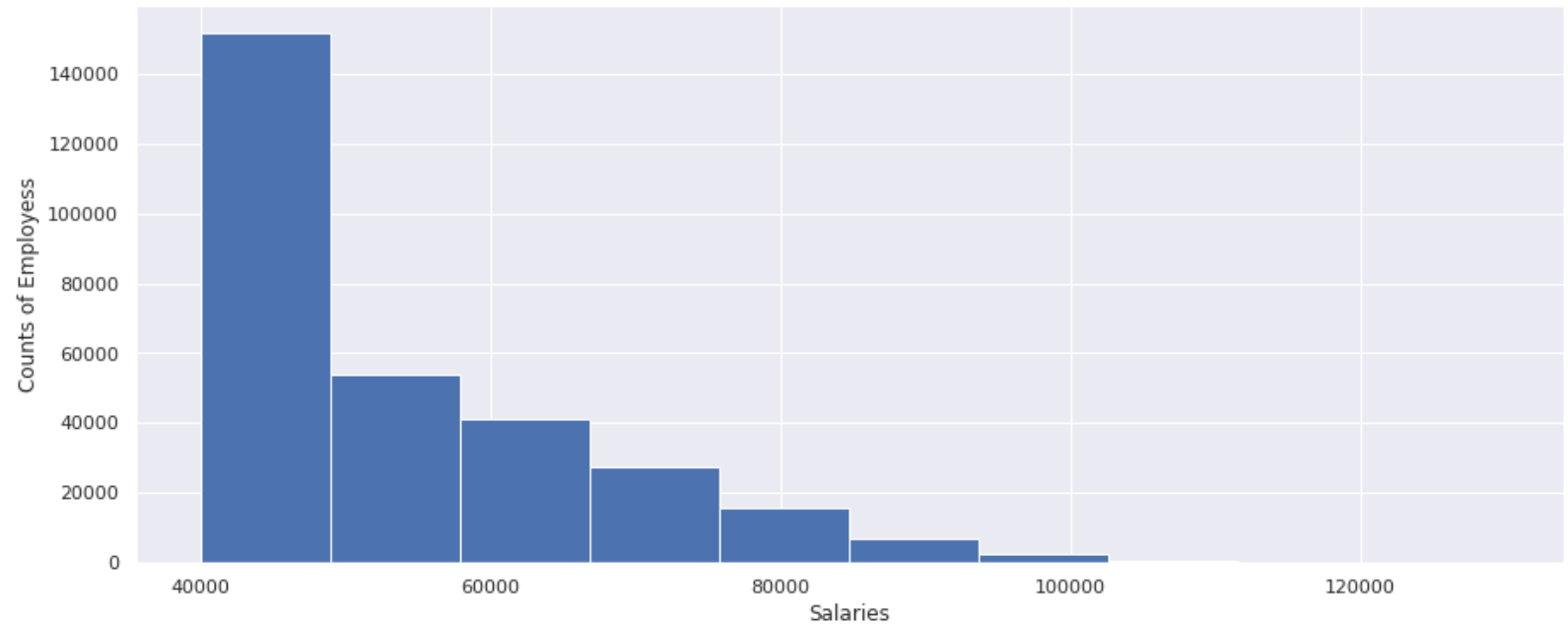


Architecture of pipeline



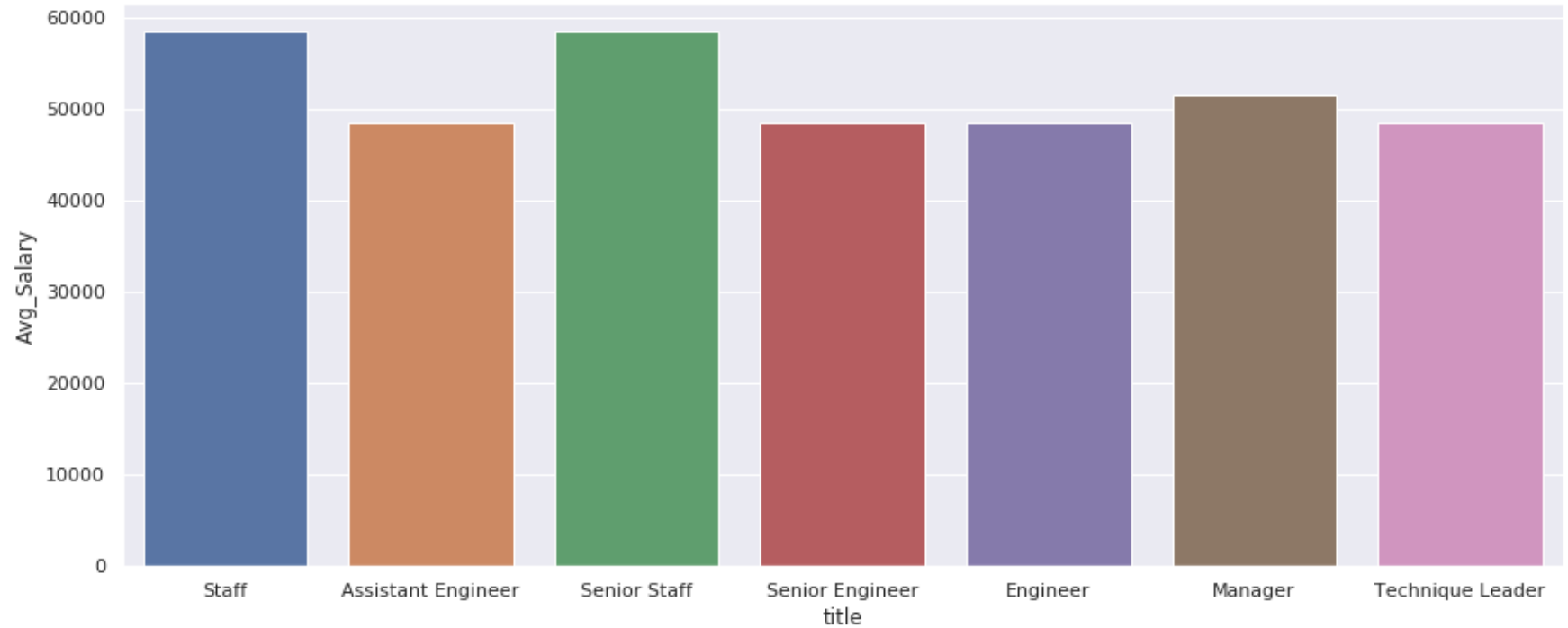
Outputs

- **Histogram to show the salary distribution among the employees**



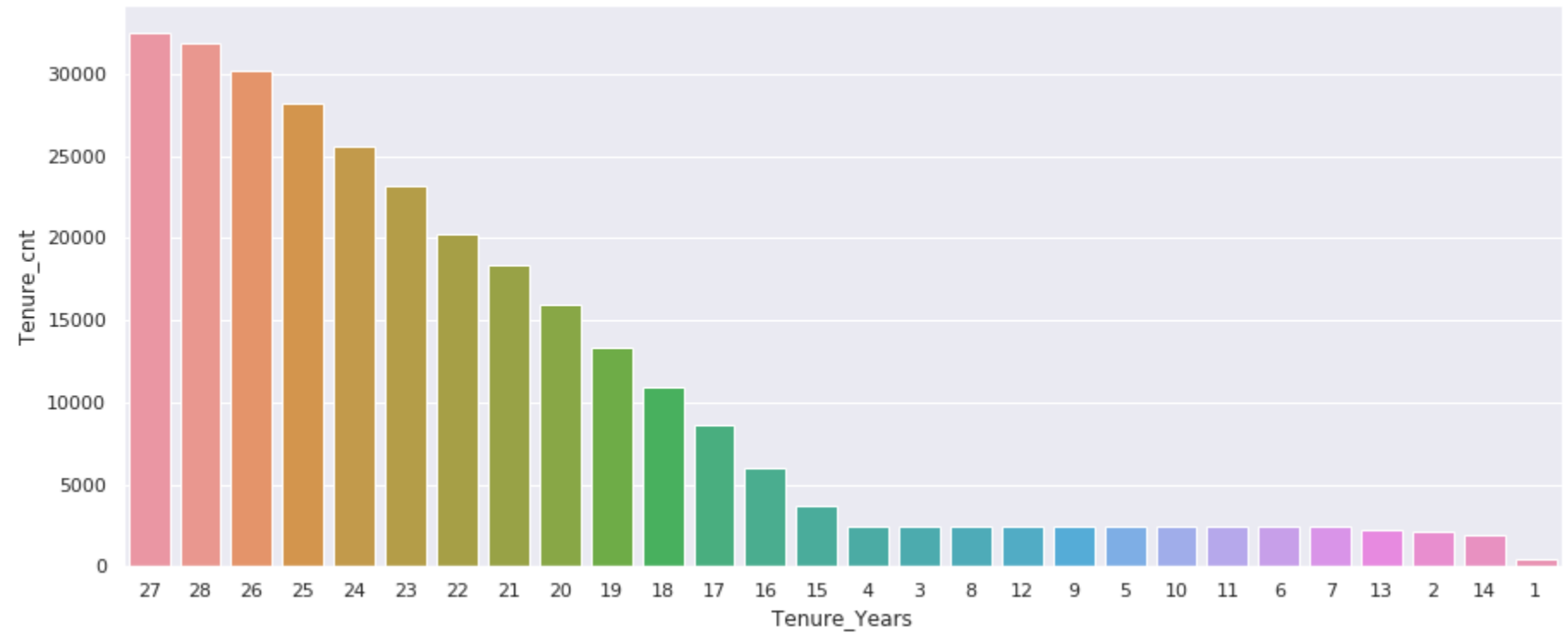
Outputs

- Bar graph to show the Average salary per title (designation)



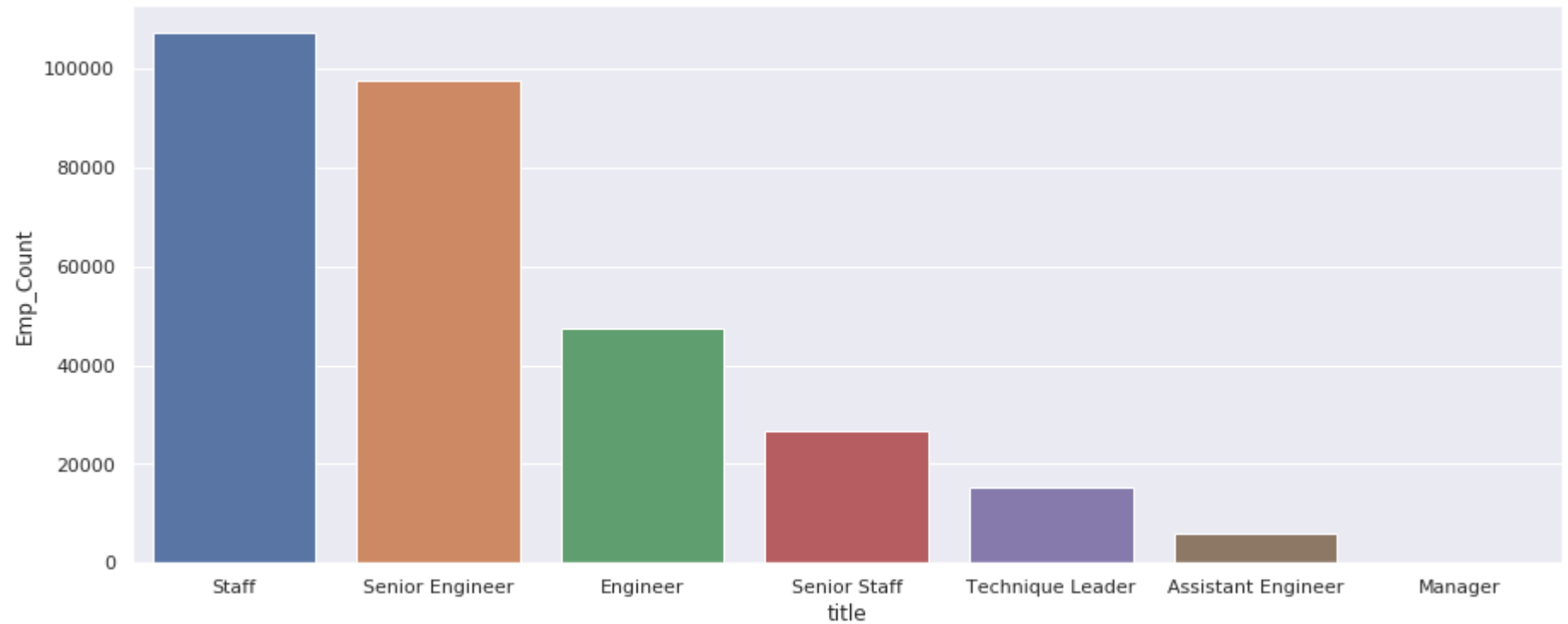
Outputs

- Calculate employee tenure & show the tenure distribution among the employees



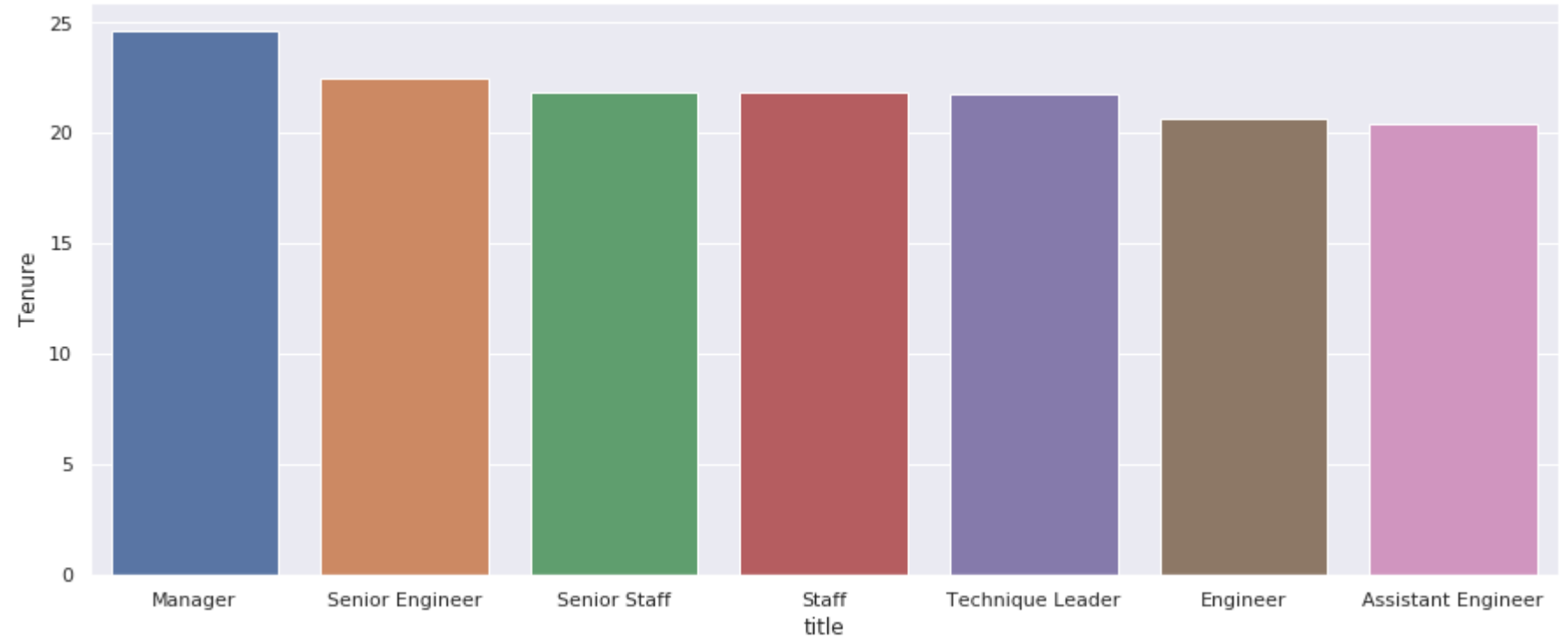
Outputs

- **Distribution of Employees across various titles**



Outputs

- **Average Tenure Distribution across Titles**



ML Model Parameters

- **Random Forest Classifier Model**

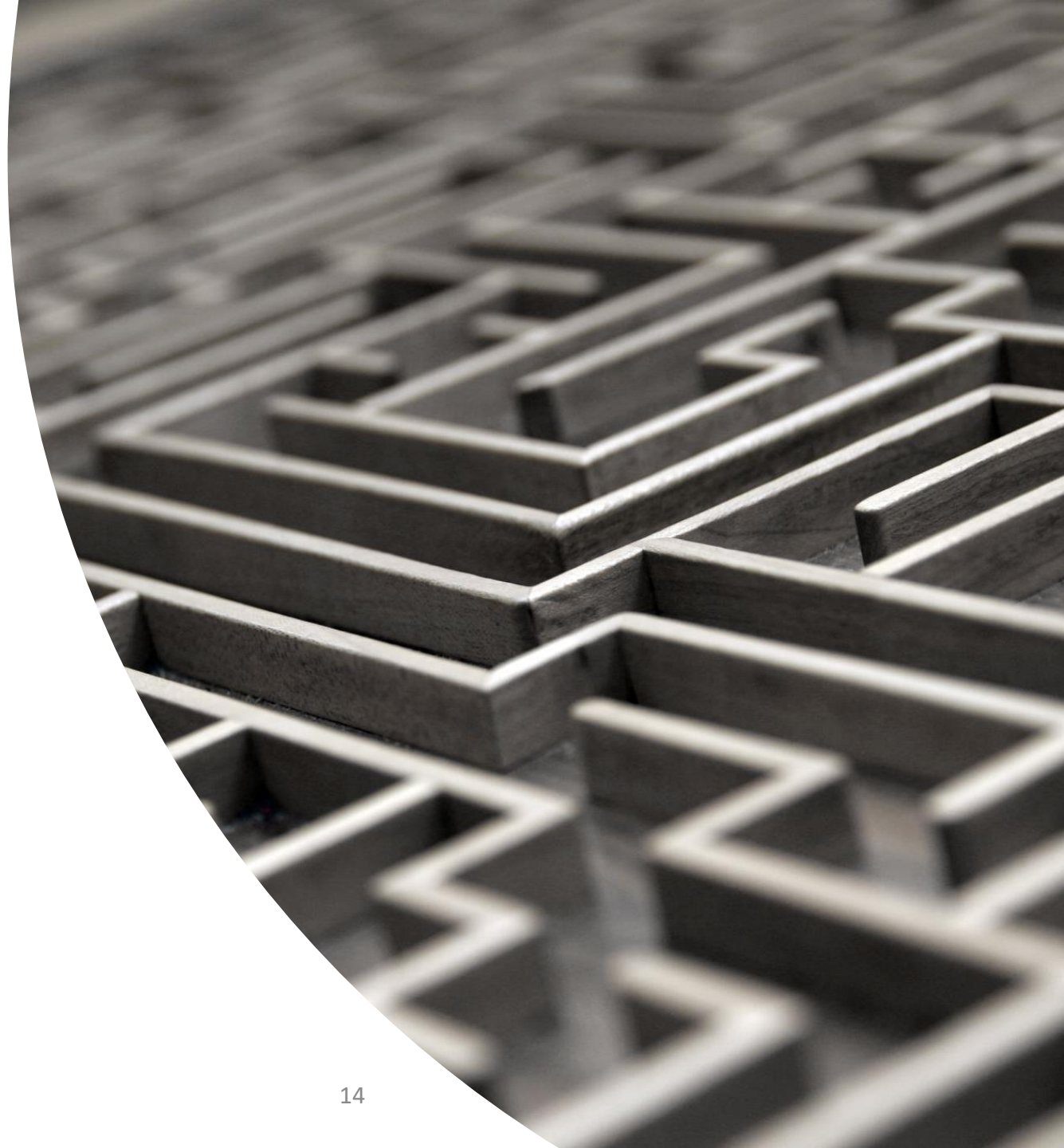
- Accuracy = 0.9980555834925744
- Error = 0.001944416507425606
- Precision = 0.9980594815941188
- Recall = 0.9980555834925743
- F1 = 0.9980471469145747

- **Logistic Regression Model**

- Accuracy = 0.9004101154357617
- Error = 0.09958988456423834
- Precision = 0.8107383759790415
- Recall = 0.9004101154357617
- F1 = 0.8532246480840692

Challenges Faced

- Importing data using SQOOP to HDFS (--m 1 argument added)
- Creating the Hive Avro Tables – Merging Data & Metadata
- Fetching the Hive tables in Spark (Jupyter Notebook Environment)
- Creating ML Pipeline as OneHotEncoding does not contain fit method in Spark 2.4.0



Next steps

- Modifying / Upgrading of Working Policies Focusing on the Employee Retention.
- Fixing the Issues related to the Appraisals / Ratings, Salaries etc. so that probability of losing a valuable employee can be reduced.
- Using the current model further analysis can be done with newly updated employee records to understand employee sentiments.

