Saurabh Salunkhe

NORTHEASTERN UNIVERSITY

Job posting analysis

**Problem Statement:**

Leveraging big data analytics to scrutinize the job postings datasets and finding relevant information like job trends, job market keywords, skills needed, new job titles etc. To get some good analytics, we can use many big data technologies like Hadoop native Map Reduce using java or higher-level frameworks like Hive and Pig. Further, data visualization can give us unique insights since we can easily find for trends by interactive visualizations rather than data formats like csv and tsv.

**Summary and Analysis**:

Having gone through the job finding process for internship, I always thought it would be cool if we could get Realtime detailed analysis of all the job market, postings, salary and trending technologies. I decided to perform big data analytics on the job posting datasets taken from Kaggle.

**Native Map Reduce using Java**

1. Median and Standard Deviation of various job categories: To get an estimate of the average job salary and spread of the salary range
2. Distinct job roles: All possible and unique job roles available for you. We can come across new job roles and further can explore that domain without wasting time
3. Count of Expired vs non-expired jobs: Getting a count of all the jobs which has expired vs jobs which has not expired
4. Pattern matching of skills (Eg. Java) keywords using distributed grep: Customizing our skillsets and getting all those job posts, which has those skills in it
5. Job roles count: At times, we want to see which job roles are repeating the most. This can give us insight about the most demanding job roles and needed skillsets.
6. Partition by all the months: Categorizing job by months can give us insight like which month has the highest opening. Which company hires in a particular month and accordingly we can plan beforehand which month to invest our time in for job search
7. Number of Jobs count in a month: Data visualization of jobs categorized by month and putting those count in Pie Chart to get insights about

**Hive:**

1. Top ten recent jobs: Sorting the jobs by postdate and thereby retrieving most recent jobs if not applied for already
2. Getting a head count of all the jobs grouped by shift, location and other parameters: Grouping jobs by various categories like shift, job type and job location
3. Getting all the jobs counts grouped by a city thereby getting an idea which city has how many jobs

**Hive Snippets:**

create table if not exists dice\_jobs(advertiserURL STRING, company STRING, job\_status STRING, job\_description STRING, jobID STRING, job\_address STRING, job\_title STRING, post\_date STRING, shift STRING, site\_name STRING, skill STRING, uniq\_id STRING ) row format delimited fields terminated BY ',' tblproperties("skip.header.line.count"="1");

load data local inpath '/home/saurabhsalunkhe/Desktop/ADBMS\_Project\_Data/dice\_com-job\_us\_sample.csv' overwrite into table dice\_jobs;

Analysis 1: Get the latest jobs by using order by clause

Query 1: select \* from dice\_jobs order by post\_date limit 10;

Analysis 2: Getting the head count by grouping by all the shifts

Query 2: SELECT shift,count(\*) FROM dice\_jobs GROUP BY shift limit 5;

Analysis 3: Getting all the jobs counts grouped by a particular city thereby getting an idea which city has how many jobs

Query 3: SELECT location, count (\*) FROM job\_location\_table GROUP BY location limit 5;

**Pig Analytics:**

Scenario: I want all the jobs in a particular city and with a particular job role.

Analysis: Getting all jobs by grouping multiple columns

Query:

1. jobData = LOAD '/home/saurabhsalunkhe/Desktop/ADBMS\_Project\_Data/PigData.csv' USING PigStorage(',') AS (location:chararray,state:chararray,jobtype:chararray,company:chararray);
2. groupdouble = group jobData by location,jobtype;
3. Dump groupdouble;

**Data Visualization using D3:**

1. Bubble chart of all the hot job keywords in the job industry
2. Pie charts of jobs posting counts categorized by months

**Spark Analysis:**

1. Real time streaming analysis of the job posts by accessing REST API (Simulated using a bunch of excel files in this scenario)
2. Performing text analytics on huge datasets and finding the trending topics using text mining, TFIDF

**Charts:**



