Most Valued Data Science Skills

Team : The DataMiners

[Ramnivas Singh, Deepak Sharma, Tage Singh, Richard Zheng, Matthew Lucich]

Dated : 03/19/2021

What are Most Valued Data Science Skills ?

Summary

- Data science primarily uses a combination of analytics and problem-solving skills
- Data science methodologies are used extract knowledge and other insights from the data
- The role of Data science is integral in many fields, key business are taking advantage of it
- Purpose of this project is to identify the most valued data science skills, as a team, we will analyze the data set to identify some important skills
- We have chosen a dataset to discover what are the most valued skills used in Data Science while working in a virtual team
- Team has applied practical collaboration, knowledge sharing, data analysis and problem solving skills on the dataset
- Team has used industry tools such as Google Trends, AWS RDS, S3, R Studio for data analysis & validations

Project Team – 'The DataMiners'

	Role	Responsibilities
Ramnivas Singh	Data & Analytics Lead	▪ Data science & Analytics execution of task with the team ▪ Catalyst to resolve impediments which arises during this project
Deepak Sharma	Data Scientist	▪ Understand the challenges, offer the solutions using data analysis ▪ Data Transformation, cleansing and visualization expert
Richard Zhou	Data Scientist	▪ Data sourcing, cleansing, modeling of structured data ▪ Data Transformation, visualization and predictive analysis
Tage Singh	Data Architect	▪ Data Architecture, ER Diagram and data management ▪ Cloud enablement, Integrations and Data Security
Matthew Lucich	Data Modeler & Statistician	▪ Understand and translate analysis questions into data models ▪ Insights from the data set, create data strategies for project

Tools & Technologies

These tools and technologies are used throughout the project to accomplish key aspects of the project

Team Communication

- Slack
- Zoom
- Outlook
- Phone / Text

Code Sharing & Quality

- GitHub
- AWS RDS
- code-inspector

Project Documentation

- MS Excel
- MS PowerPoint
- Rpubs
- Workbench Reverse Engineering

Data Load & Cleansing

- .CSV import
- R Data Import
- AWS S3

Data Mining & Analysis

- Google Trends
- AWS RDS
- dployr
- tidyr

Data Source & Loads

Data to Collect

•Job listing data with attributes including: job title, job description, salary, company, location

•Collected in last 2-3 years

Data Location

•Web scraped data from Indeed job descriptions is hosted on Kaggle

•Data location: https://www.kaggle.com/elroyggj/indeed-dataset-data-scientistanalystengineer?select=indeed_job_dataset.csv

Load Data

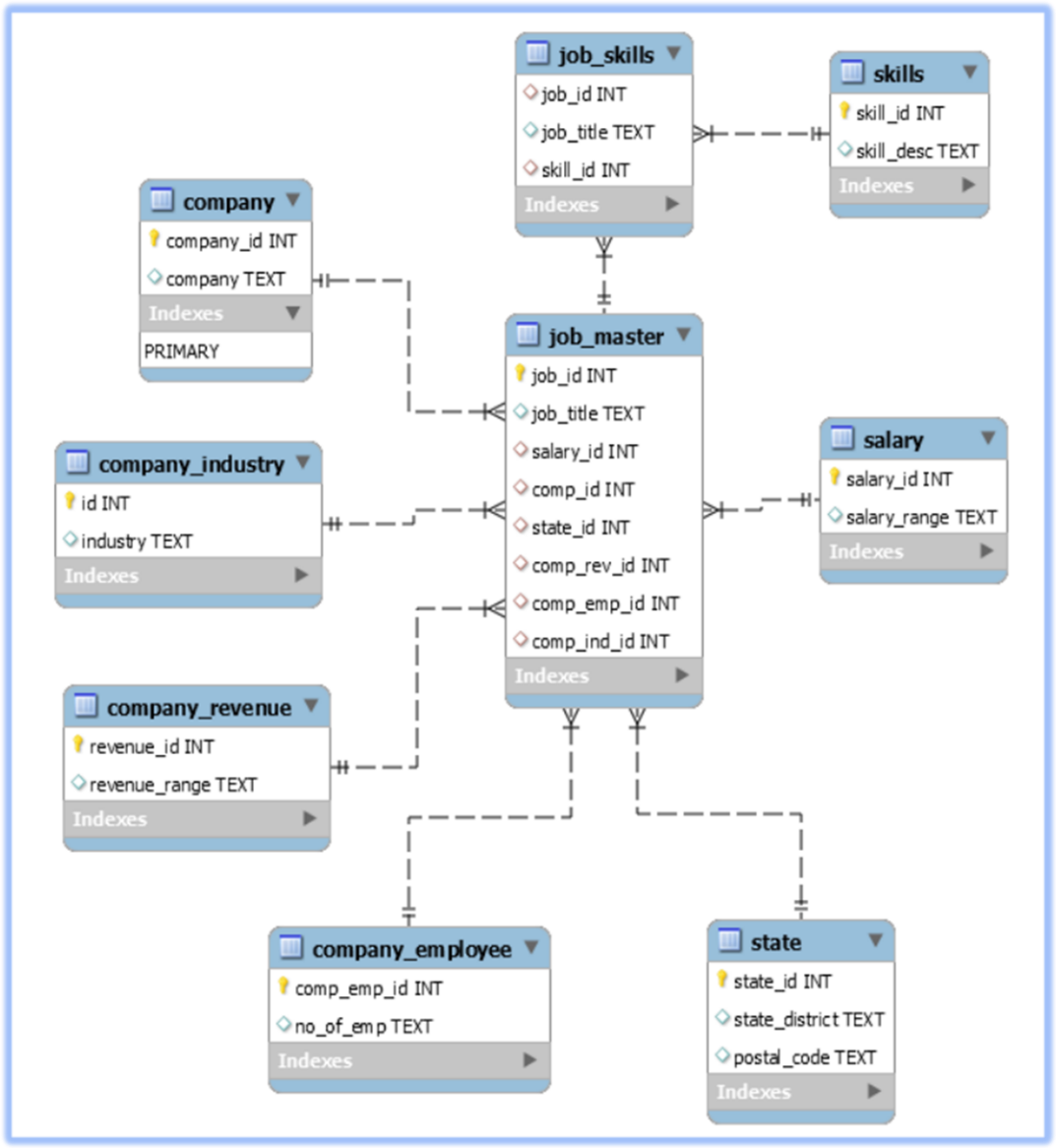
• Load Kaggle CSV into Amazon RDS database via the "LOAD DATA INFILE" statement
    - Specify appropriate values for "FIELDS TERMINATED BY", "ENCLOSED BY", "LINES TERMINATED BY" and other statements related to data formatting

• Once inserted into RDS, load data into project's R markdown file by connecting to AWS database through the R package: RMySQL
    - Utilize dbplyr's in_schema function to access tables using non-default schema
    - Utilize base R's as.data.frame function to coerce results to dataframe

• Credentials will be read from our environment variables for improved security

Database Design

Entity Relationship (ER) Diagram



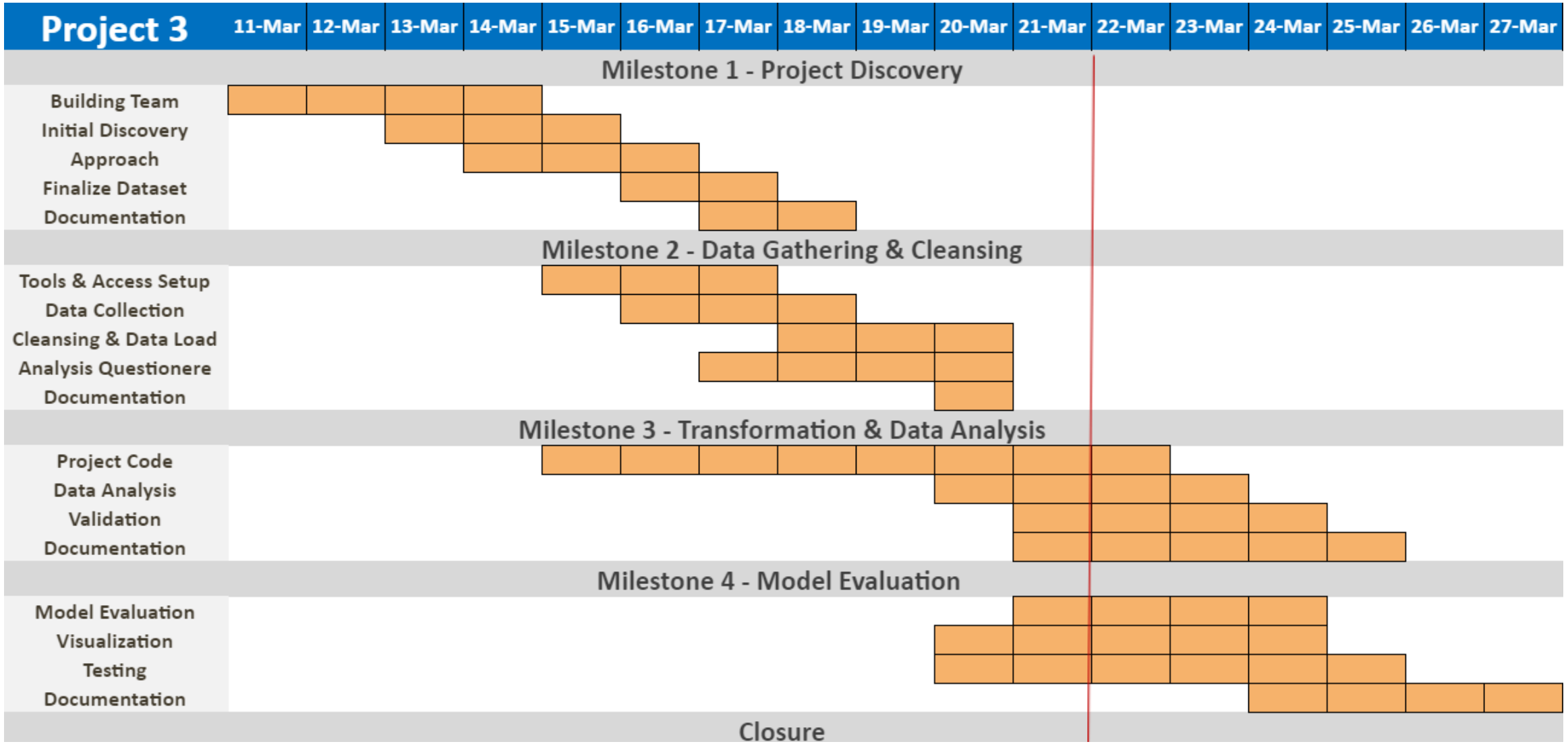
Database Objects & Provider

job_skills : This table keep required job skills for a job posting
skills : This table is used to retain a master list of the skills
salary : This table keep salary mentioned on a job posting
state : State for which this job is posted
job_master : This is a key table to retain job details.
This key table for data analysis
company_employee : To retain employee count of job posting company
company_revenue : To retain total revenue of job posting company
company_industry : To keep business industry of job posting company
company_revenue: Name of the company posting data science job

view_d607_p3_all_recs: view to return complete data normalized data

Provider : Amazon Web Services (AWS) RDS
Database Host : data607-project3.cbs1lxtno2zh.us-east-2.rds.amazonaws.com
Database Port : 3306
Database Security : IP Based Access, VPC security groups
Password Management : laresbernardo/lares package

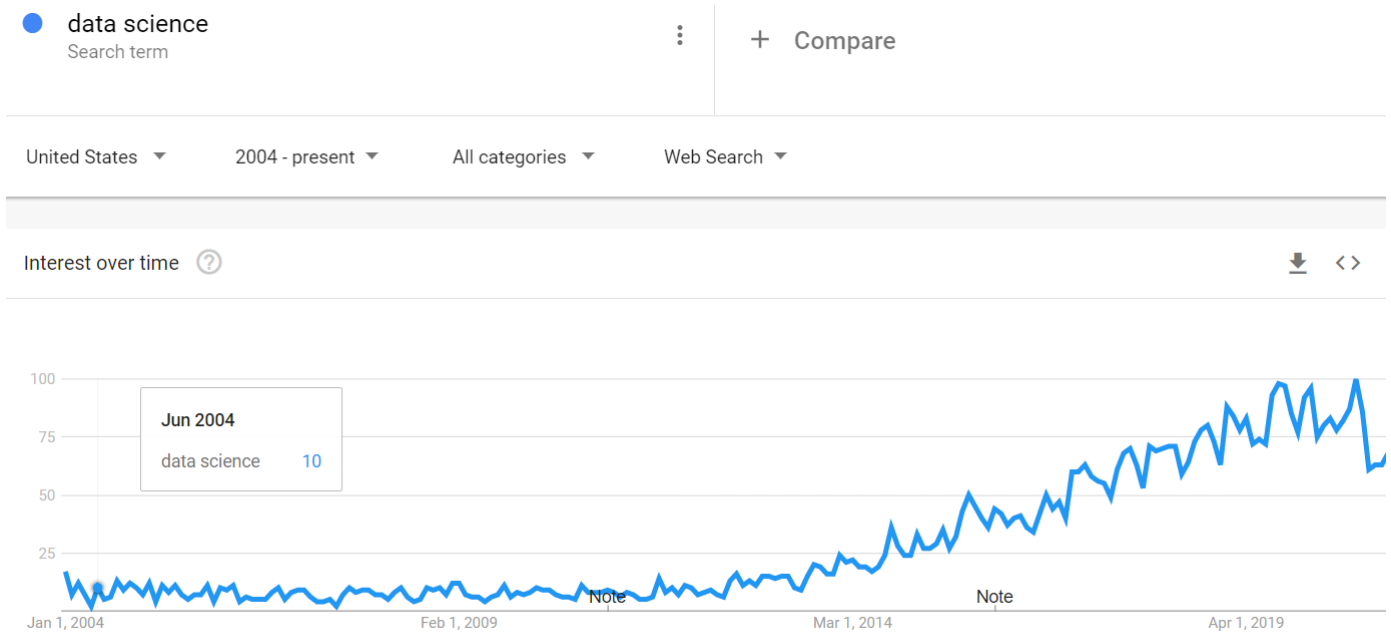
Approach & Execution

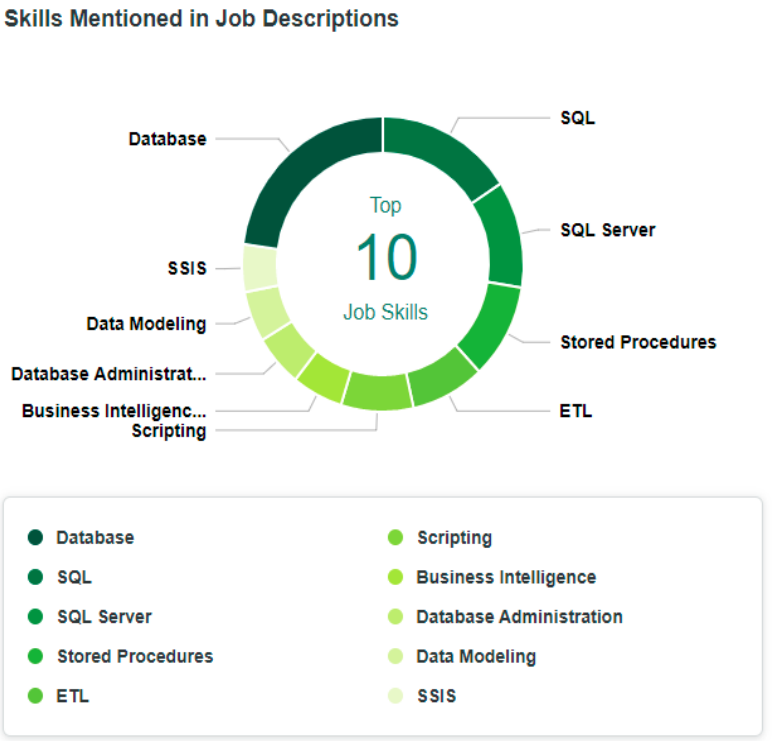


Team to deliver….

Data Analysis

- What are soft Skills
- Are Data Science Soft Skills unique to this type of work
- What about Data Science Managers, are their soft skills requirement unique or are they similar to other managerial soft skill sets in tech?
- What are important skills for Data Science
- Do important skills vary by location, experience, salary?
- How has the interest for skills changed over time?





Trends & modeling

Google Trends

•For top skills, filtered on various attributes (e.g. salary, location, etc.), Google Trends will be scraped

•Data will be analyzed to determine:

- What skills and technologies have increased in popularity over time?
- What skills and technologies have decreased in popularity over time?
- What skills and technologies have remained stable in popularity over time?

Modeling

•Modeling section will include 1-2 approaches to be determined, which may include:

- Regression
  - Predict salary based on:
    - Skills
    - Years of experience
    - Location
- Clustering
  - Is interest in skills and technologies clusterable by:
    - Years experience
    - Top companies (by revenue)
    - Top industries (by count)