Russell Davis
Walter Gerych
Jingwen Sun
Jie Fu

MIS 584 FINAL PROJECT REPORT

Business Intelligence Solution for Indeed

1. Executive Summary

In our time observing Indeed's inner and outer workings, we have seen many areas of strength and effective operating procedures. The strength of Indeed lies primarily in the technical acumen of its employees, and we are also impressed by the level of organizational support for business intelligence efforts within the organization. After some investigation, we have two proposals intended to further refine Indeed's practices and make productive use of their existing business intelligence infrastructure.

Our first proposal is intended to enhance current practices for tracking project completion and individual contributions of programmers in the Engineering department. The current approach makes use of tabular data with whatever practices are desired by each management team; there are no universal tools provided to ensure that all teams are focused on the same KPIs and observing them in the same manner. Our report contains our prototype for an interactive dashboard made with Tableau that will help to unify and coordinate Indeed's approach to project management. The prototype is sufficiently functional in its current state to provide a proof of concept. However, we lacked full access to Indeed's internal data while creating it, leading to significant problems with completeness of our data. With the template we provide, managers at Indeed with full data access should avoid this obstacle completely.

This dashboard, which will primarily provide benefits at the operational level, provides a summary of Jira commits associated with each active project at Indeed and allows project managers to monitor the progress of their team and the state of their project. This is done by providing project managers with KPIs such as Percent of Open Jobs and Change in Commit Rate, as well as allowing the project manager to quickly access the breakdown of what kind of commits are being made to their project and pick out which team members providing the most commits. Additionally, we demonstrate the benefits of using analytics by clustering employees by commit type and by number of projects worked on. By clustering by commit type, we identify four general types of commit behavior. This is beneficial for when teams are being constructed for a new project, as by ensuring that each cluster is represented in the new team the project manager can ensure that his team will have the necessary skills to produce a successful project. By clustering by number of projects worked on, we provide upper management with a way of quickly assessing the experience level of employees whose history they may be unfamiliar with.

Our second proposal is to refine the manner in which search results are presented to consumers. The current presentation lacks aesthetic refinement and, moreover, contains few visualizations. Visualizations are a key method of rapidly presenting a large amount of data in an easily understood way, making it an obvious point of improvement.   Specifically, we intend to rapidly convey the locations and salary of job openings in the United States both locally and based on national average, with easy comparison between different job titles.

One problem we experienced was that Indeed did not provide an easy means to acquire the information on average salary by city and profession, necessitating the need for us to develop a web scraper to procure this information. This constituted a significant time investment. We anticipate that Indeed's internal processes facilitate rapid procurement of this data, as the data warehouses should mean that web-scraping should not be necessary. However, if this is not the case, we strongly recommend that this capacity is developed.

The external dashboard we developed using the data we obtained from the webscraper allows users to compare the average salary and job demand between cities of their choosing. Additionally, this dashboard allows one to filter by one or more job titles and compare the salaries and demand for these job titles by city, as well as the relative demand for entry, mid, and high level jobs for each city/job title combination. Moreover, this information is presented to the user with simple visualizations that can be quickly understood.

2.Introduction

2.1. Client Company Background

Indeed, which hosts an employment-related search engine at Indeed.com, provides free resources to connect potential employers to potential employees. Indeed competes with well-known corporations such as LinkedIn, CareerBuilder, Monster, and ZipRecruiter, all of which provide similar job-placement services. Indeed is the most trafficked job site in the world as of March 2017, and as of September 2016 received more than 200 million unique views on a monthly basis ("About Indeed", 2018).

Indeed is headquartered in the United States, with primary offices in Austin, Texas and Stamford, Connecticut. It has 19 other offices located in 13 different countries and a total of over 6,100 employees. Indeed is comprised of 12 distinct departments. These include standard corporate departments, such as Finance, Legal, and Human Resources, and departments common to technology companies such as Software Engineering and an Technology Services department, which is made up in large part of information technology roles. Indeed also has departments specific to their role as a search engine, including Search Quality and User Experience.

As indicated by its large footprint and heavy traffic, Indeed hosts jobs on a global scale, with listings for positions in over 60 countries across all 6 populated continents. Indeed's core offering is their vertical search engine oriented around job search and aggregated job listings, but it also provides ready access to company reviews, industry- and company-level salary information, and individual resumes. This wealth of information about industries and employment opportunities is useful to both employers and job seekers, and also provides Indeed with a tremendous amount of data to analyze. As a result, a significant portion of their employees are familiar with software engineering and SQL-style querying.

The company-wide comfort with technology also means that most necessary software is intended for use across the company, instead of being exclusive only to departments that possess the requisite expertise. These applications are most often provided by Indeed's Software Engineering department, which ensures that the necessary tools are developed and maintained to support business operations. A large part of our project will be focused on this department, although the Business Intelligence solutions we present are likely to be beneficial for other departments as well. We also propose to employ the user-friendly Tableau software to create outward-facing data visualizations in a manner that will be discussed in more detail later in this report.

2.2. Business Intelligence at Indeed

As mentioned above, Indeed's primary offering is their vertical search engine oriented around job search, and their business model is accordingly is centered around this role. As such, data is their lifeblood, and they have made use of large-scale data warehouses for much of their existence. These warehouses are hosted by Amazon Web Services to ensure cost-efficient storage and accessibility on a global scale.

Indeed does not make use of an ERP to support operations, relying instead on a Customer-Resource Management tool powered by Salesforce, and a variety of other tools to manage internal data. Various departments make use of big data tools such as Hadoop and Apache Hive, as well as the aforementioned cloud services, streaming data pipelines such as Spark Streaming, and various sorts of specialized proprietary software depending on the department's needs.

In addition to these tools, the primary mechanism for Indeed's data sharing and analysis is a large-scale analytics platform developed in-house called Imhotep ("Imhotep - Data Analytics Platform by Indeed", n.d.). This system is used across the company and uses a query language similar to SQL to facilitate ad hoc querying and data aggregation. Imhotep is used to access a wide variety of data that is pertinent to all engineering and product-related departments within Indeed. This includes data about website use meta-data (e.g. time spent on web pages, mouse-click tracking, and data pertaining to website navigation) as well as information about the queries themselves and other data hosted on Indeed (e.g. search keywords, geography-based search data aggregation, salaries, and other data about the job listings themselves).

Imhotep also supports decision trees and machine learning algorithms, visualizations, and the development of scripts, dashboards, and applications tailored to specific departmental needs. It currently supports 6 analytic web apps, 5 dashboards, 10 programming/scripting shells, 6 monitoring apps of production software, and than 40 internal clients (Indeed Engineering, 2014). The knowledge prerequisites for making effective use of Imhotep are significant, but the variety of possible tasks that it enables make it a valuable resource for managers and the existing Business Intelligence teams within Indeed.

Indeed's BI teams are primarily located within the Technology Services department, working alongside IT personnel and database management teams. Some teams are focused on analysis, others are dedicated to support specific high-priority teams, and individual BI personnel are embedded in other groups to provide a unique viewpoint and expertise. These teams benefit significantly from the volume of data available at Indeed, which allows them to access a plethora of information regarding customer interests, employer requirements, industry needs, and a variety of other topics.

2.3.Evaluation Using Eckersley's Business Maturity Model

In order to quantify the level of maturity achieved by corporate BI efforts, Wayne Eckerson (2011) developed a 5-stage Business Intelligence Maturity Model. These maturity levels range from companies with fledgling BI practices to those that offer highly advanced BI services. Early hallmarks of maturity include internal financial support and sponsorship of BI as a concept and minimization of redundancies in data handling. Later stages of maturity are characterized by ready user access to centralized high-quality data, cross-departmental synergy with respect to data warehousing and terminology, and the ability to maintain BI cost-effectiveness and relevance as the program continues to grow.

Indeed has shown a willingness to embrace BI, as evidenced by support of independent teams and cross-departmental integration of BI experts into other teams. Indeed's holistic reliance on data has naturally led it to effective warehousing, sustainable delivery of near-real-time data, and other hallmarks of data management mastery. BI techniques and approaches are applied in public-facing venues such as the Indeed Hiring Lab, which is comprised of economists and researchers and "produces research on global labor-market topics using Indeed’s proprietary data and publicly available sources" ("About Indeed Hiring Lab", 2018). Indeed also has Imhotep features that are explicitly designed to provide visualizations for employment trends for various industries, as well as user searches and other compilations that can help with generating dashboards and BI products.

However, we find that Indeed falls short of Eckerson's highest level of maturity. This level, which he refers to as "Sage", calls for the organization to "turn the BI resource inside out and [make] it available to customers and suppliers". Here, the inherent complexity of Imhotep works against Indeed. While Indeed has made Imhotep open source and published useful guides, the available interface is still beyond any user that lacks a solid foundation in SQL and experience with data aggregation. Also, the results produced by Imhotep are generally focused on functionality rather than aesthetics. This can limit its usefulness for public-facing contexts for both Indeed and potential outside users where appearance is a relevant component of a product's appeal.

Additionally, Eckersley also calls for embedding BI deeply "into the business processes that drive the company". We see that the flexibility of Imhotep lends itself so readily to novel uses that there is a lack of normalized company-wide products or processes. We conjecture that this is also partially due to the complexity of Imhotep, as some of the more business-oriented departments (e.g. Human Resources) may struggle to be effective with applications that a technically-oriented department uses. This lack of a coherent approach prevents BI from being fully embedded. Together with the previous shortcoming, we suggest that Indeed should be classified as having Adult maturity, trending towards Sage-level maturity.

3. The Proposed Business Intelligence Solution

As hinted at above, Business Intelligence at Indeed is broadly responsible for the paradigms and processes needed to parse data and transform it into an actionable state. Indeed has already had a great BI implementation in the company. However, it's not good enough. It still has disadvantages. The Imhotep is customized but not intuitive. It also has complexity. Our project is to use a better visualization and easier to use tools - Tableau, to develop two dashboards, one is internal and the other one is external. The internal dashboard will mainly focus on the software engineering department. It will help the project manager to keep better track to the engineers' performances. The external dashboard is customer-oriented. It will generate more useful information than that Indeed currently can provide to customers.

3.1. BI Framework

The four components are people, process, management, and governance. As an external dashboard is not really for improving the efficiency of Indeed, the BI framework of our project would be discussed in details only for the internal dashboard.

3.1.2. People

The people mainly represents the software engineers in Indeed. They have a lot of projects to do. The data is generated from their daily works and the information extracted from the data is used for improving their works. The value of this component is to provide the essential data needed for the Indeed BI solution.

3.1.2. Management

The management in the BI frameworks means managing the data. After the collection of data, the data should be transformed into a proper format and be loaded into the data warehouse. This process contains cleaning data and aggregation, which can ensure the data quality before the data being used. The value of this component is to keep the data cleanliness and accuracy so it would avoid the situation that the analysis is interrupted due to the data quality issue.

3.1.3. Governance

After loading the data to the data warehouse, our team will take the needed data and put them in the small data-mart. Indeed should take charge of the updating of the data warehouse. There will be new data every day, Indeed should put them all in the warehouse to keep the data accuracy. Our team should update the data mart simultaneously as well.

3.1.4. Process

After the data work, we will then use the data in the data marts to provide an internal data visualization for Indeed to evaluate the performance of software engineers. Also, through Tableau, our team will provide Indeed with additional clustering analysis to show that Tableau is not only a visualization tool, but it can also do data analysis as well. And this clustering analysis can improve the decision-making process from other perspectives. This component is the real step when we deliver the useful information and value to Indeed.

3.2. Data Source

Our product needs two kinds of data. The first one is for the internal, which can be called operational data source. The other one is for the external, which can be called external data source.

3.2.1. Operational Data Source (Internal)

Operational data generally means the operation data which can be collected from daily work. For our product, the operational data is the same as general. We need the data generated from the software engineers' daily works. For Indeed, there is a platform called Jira. This is a platform where the software engineers can get their mission and submit the commits of their projects on hand. The Jira contains the software engineers' names, projects names, and projects status etc. Therefore, for the operational data, our data source is only Jira because it has all the data might be useful for our product.

3.2.2. External Data Source

External data source generally means the data from outside of the company. However, in Indeed, it's different because the product of Indeed is data. The external data source for our product is the job posting information from the official Indeed website. The data from the website is generated by Indeed itself because it's a platform at where companies are willing to post a job description. Therefore, the external data source is Indeed itself but it's not internal data. Our team member wrote a web scraper to transform the job posting to excel dataset. The dataset contains the job title, the salary of this job, city, top 5 employees, level of jobs and types of jobs(internship or full-time etc.). The code would be shown in the Appendix.

3.3. Data Model

Indeed has already had a big-scale data warehouse. However, for our product, what we need are small data marts for each dashboard. In order to get a better understanding of how two data marts can integrate useful information, we should create two simple data models to know the relationships between different dimensions with the facts, which are the actual information we want to know.

3.3.1. Star Schema - Internal

Figure 1 is the Star Schema for internal data

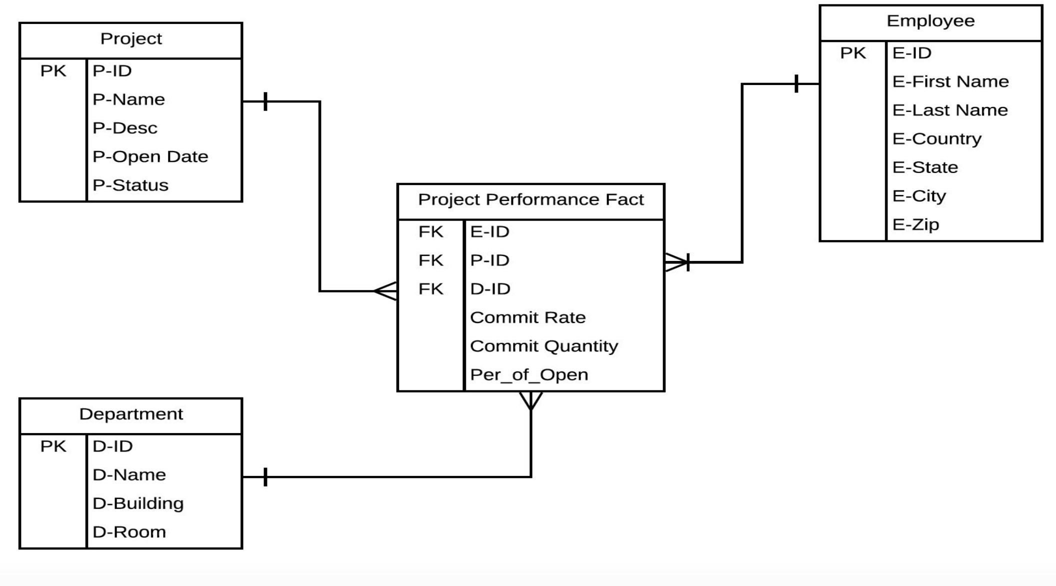


Figure1. Internal Star Schema

Facts: Project Performance Fact

Dimensions: Project, Department, Employee

This model is for the internal data. The employee information is the information of the software engineers. With the employee information, project information and department information, we can evaluate the metrics about project commit rate of each employee, the total commit quantities and the percentage of open commits, etc..

3.3.2. Star Schema – External

Figure 2 is the Star Schema for external data

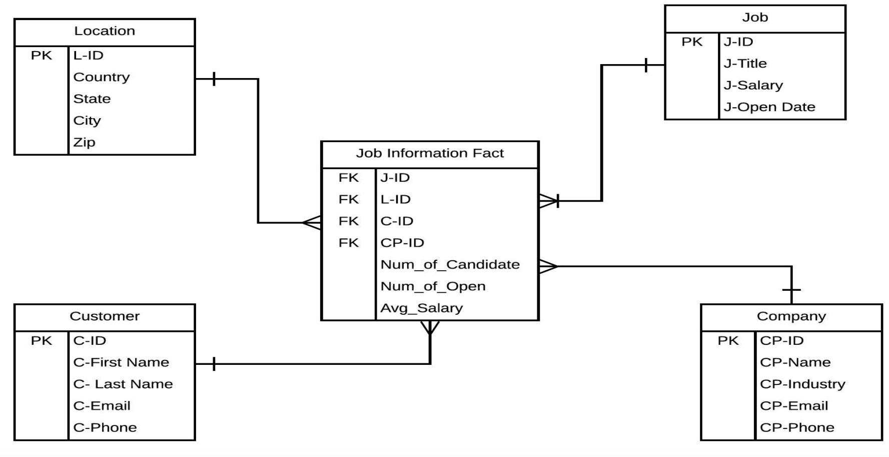


Figure 2. External Star Schema

Fact: Job Information Fact

Dimensions: Location, Job, Company, Customer

This model is for the external data, The job is the job information that has been put on the website. The customers are the individual users who have uploaded their personal information on Indeed. Companies are the companies that post a job position on Indeed. With this information, we can evaluate the metrics about the number of candidates, the number of demands, and the average salary of each job in each city, etc..

3.3. Monitoring, Data Analytics & Data Visualization Functionalities

After creating data marts and data models, we can now use data to do the visualization and other functionalities. We used Tableau to develop two dashboards for Indeed. The first one is the Operational Project Management Dashboard and the other one is the External Customer job searching Dashboard. Besides these two dashboards, we also make clustering analysis by Tableau to show Indeed that Tableau is not just a visualization tool, it can also provide advanced analysis. And the results from the advanced analysis can help Indeed as well.

With the operational project management dashboard, the project managers in Indeed can keep better track to the project process. They can see the project directly from the dashboard and then can know whether this project needs attention. Also, they can better evaluate the software engineers' performances. The key performance indicators in this dashboards are total commits, team sizes, change in commit rate and percentage of open jobs. "Open" means a work that has not been done. Total commits represents the total quantities of commits that a software engineer have submitted. With this KPI, the project manager can see the quantities of commits of both a software engineer and a project. Team size is the number of software engineers in the projects. It can show how many engineers are working on one project at the same time. With this information, the manager can have thoughts about the team size and can decide whether he needs to assign more people to this project. Change in commit rate is based on quarters. The change rate shows whether the commit rate of this project increases or decreases compared to previous quarter. It works for worker as well as it enables the manager to see the commit change rate of each employees. If the manager sees from the dashboard that the commit rate decreases sharply, he might want to take actions about it. The percentage of open job means how many projects are now have a open status, in other words, not be finished yet. The alert level of the percentage rate is 20%. If any project or software engineer has open percentage higher than 20%, it means this project needs attention also there might be some troubles on work for this engineers. With this intuitive vision of the software engineers' performances, managers can know the working pattern of each employee. Once the patterns show unusual appearance, managers can take actions, therefore it will increase the overall efficiency in the software engineering department.

With the external job searching dashboard, the users of Indeed website can get more useful information than from the website. The website only has job descriptions, but with the use of our external dashboard, customers will find average salary of the job directly, and also the hiring demands etc., which are all the factors that people will care about while looking for a job. The KPIs in this dashboard are national average salary, city average salary and hiring demand. The national salary is the average salary of each job by all cities. The city average salary is the average salary of every job in each city. The national average salary can be the baseline of salary. If the city average salary of one job is lower than the national salary, it means the salary in this city is at a low pay rate. The customer can know directly about the pay rate then can decide whether they are satisfied with that. Also, customers also can compare the average salary of one job between different cities. If the customer has several options of working location, he or she could use this KPI to see which location has a highest pay rate. That would be really helpful for finding a working location. The hiring demand represent the total quantities of job listing of one job. From this, the customers can know the competition of this job is severe or not. These functions can make users have a better experience while using Indeed. Nowadays, people use Glassdoor, LinkedIn quite often to find a job. They have similar functions as Indeed website so they are actual competitors for Indeed. With our external dashboard, people who are looking for jobs can know better about the jobs than through other platforms. Therefore, the purpose of the implementation of the external dashboard is to attract more customers to use Indeed so that Indeed can get more market share in job searching area.

Clustering analysis is done mainly for the internal dashboard. Different segments have different patterns, it can help managers to know the classification of employees and make accurate decisions to increase the work efficiency. Our team did two clustering analyses. The first one is clustered by commit type, it can identify four general types of commit behavior. This is beneficial for when teams are being constructed for a new project. Another one is clustered by number of projects worked on. This analysis can provide upper management with a way of quickly assessing the experience level of employees whose history they may be unfamiliar with.

Figure 3 is the Framework of Business Intelligence Solution for Indeed.

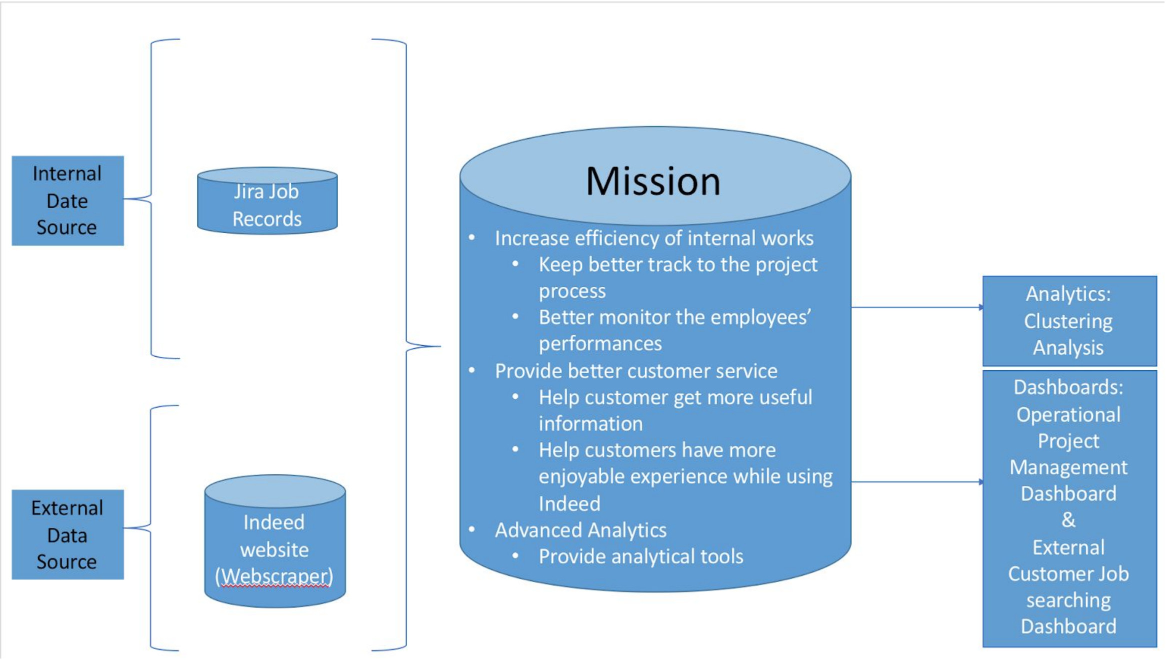

Internal Date Source
Jira Job Records
External Data Source
Indeed website (Webscraper)
Mission
Increase efficiency of internal works
Keep better track to the project process
Better monitor the employees' performances
Provide better customer service
Help customer get more useful information
Help customers have more enjoyable experience while using Indeed
Advanced Analytics
Provide analytical tools
Analytics: Clustering Analysis
Dashboards: Operational Project Management Dashboard & External Customer Job searching Dashboard

Figure 3. The Framework of Business Intelligence Solution for Indeed

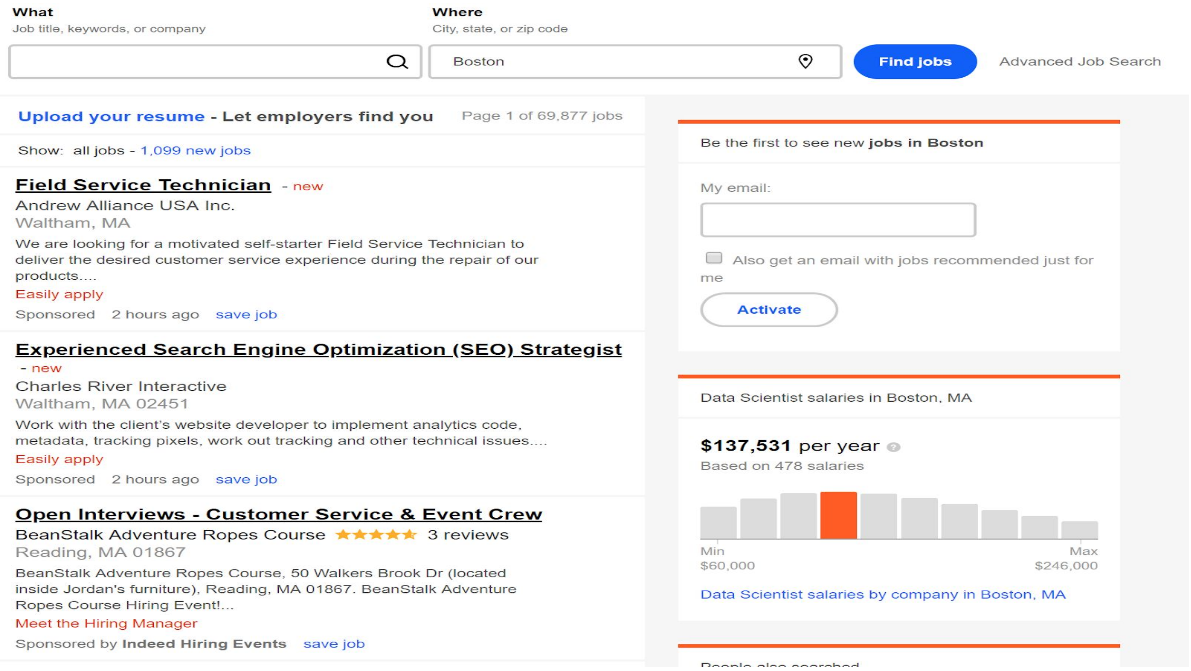
4.Use Cases & Prototypes

We have created both an operational and a customer dashboard to illustrate how our BI solutions can benefit Indeed, which we will demonstrate in detail below. The external dashboard we developed lets users compare the average salary and job demand between any number of cities and job titles. This dashboard also makes the relative demand for entry, mid, and high level jobs for each city/job title combination readily available and easy to understand with a simple visualization. The operational dashboard provides a summary of Jira commits associated with each active project at Indeed and allows project managers to monitor the progress of their team and the state of their project. It also clearly lists KPIs such as Percent of Open Jobs and Change in Commit Rate.

4.1.The Customer Dashboard

We will demonstrate how Indeed can use its data to offer its customers increased value with a prototype of a customer-facing dashboard. This dashboard allows the customers to investigate the average salary and demand of various job titles between multiple cities.

Before we present our visualization solution, let's examine how Indeed currently presents its data to its customers.



Boston

Upload your resume - Let employers find you   Page 1 of 69,877 jobs

Show: all jobs - 1,099 new jobs

Field Service Technician - new
Andrew Alliance USA Inc.
Waltham, MA
We are looking for a motivated self-starter Field Service Technician to deliver the desired customer service experience during the repair of our products....
Easily apply
Sponsored 2 hours ago save job

Experienced Search Engine Optimization (SEO) Strategist - new
Charles River Interactive
Waltham, MA 02451
Work with the client's website developer to implement analytics code, metadata, tracking pixels, work out tracking and other technical issues....
Easily apply
Sponsored 2 hours ago save job

Open Interviews - Customer Service & Event Crew
BeanStalk Adventure Ropes Course ★★★★★ 3 reviews
Reading, MA 01867
BeanStalk Adventure Ropes Course, 50 Walkers Brook Dr (located inside Jordan's furniture), Reading, MA 01867. BeanStalk Adventure Ropes Course Hiring Event!...
Meet the Hiring Manager
Sponsored by **Indeed Hiring Events** save job

Be the first to see new **jobs in Boston**

My email:

□ Also get an email with jobs recommended just for me

Data Scientist salaries in Boston, MA

$137,531 per year
Based on 478 salaries

$60,000 $246,000

Data Scientist salaries by company in Boston, MA

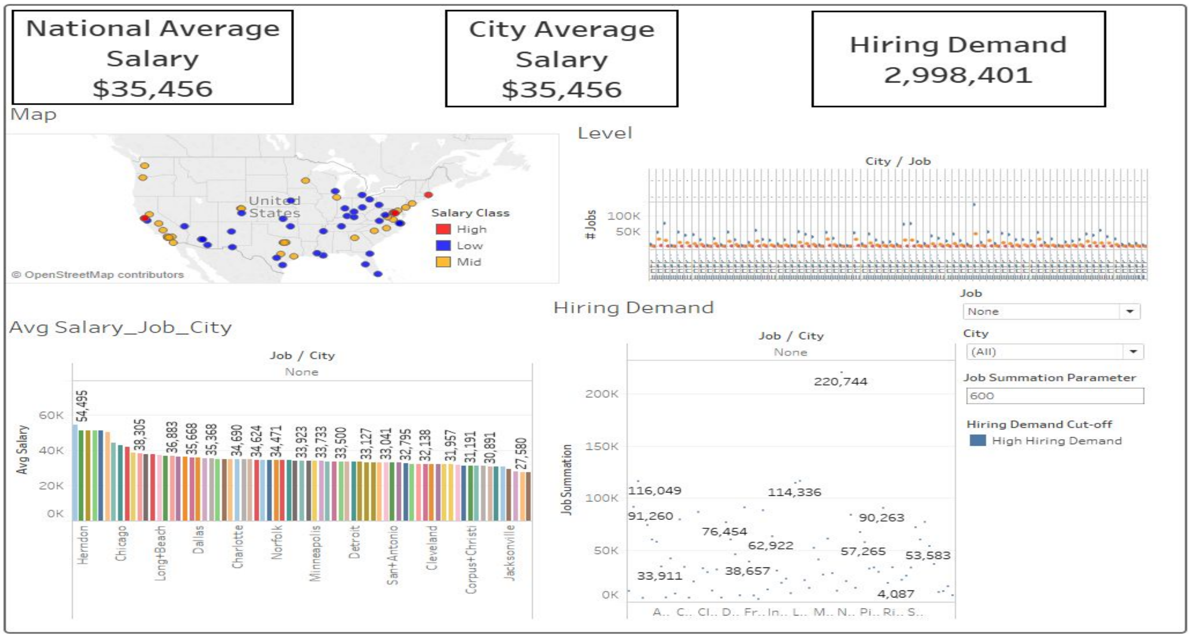
People also searched

Figure 4. Indeed website

The above figure shows the current method of presenting your data to your customers at Indeed. The user can search by job title and city, and is then shown a list of jobs matching the criteria along with a simple plot showing the mean salary of their query. This is adequate if the user is interested in searching for jobs associated with only one job title, and if they are interested in looking at only a specific city. However, comparing various job titles or the job prospects between a set of cities is not made easy. We will now demonstrate our customer dashboard, which addresses the shortcomings of Indeed's current method of presenting customers with its data.

The Customer Dashboard: Default View

We developed an external dashboard the data we obtained from a web scraper that pulled Indeed's publically available job listings. This dashboard allows users to compare the average salary and job demand between cities of their choosing. Additionally, this dashboard allows one to filter by one or more job titles and compare the salaries and demand for these job titles by city, as well as the relative demand for entry, mid, and high level jobs for each city/job title combination. Moreover, this information is presented to the user with simple visualizations that can be quickly understood.



National Average Salary $35,456

City Average Salary $35,456

Hiring Demand 2,998,401

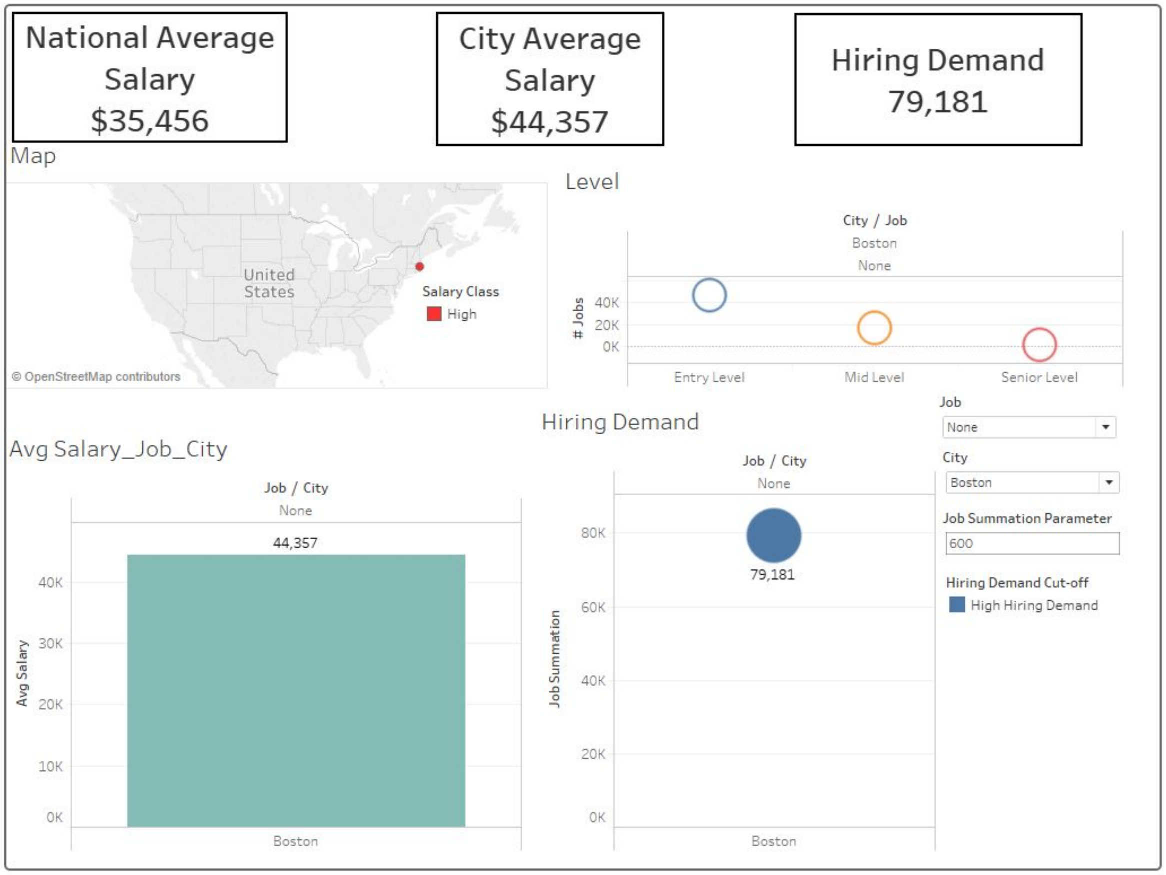
Job: None

City: (All)

Job Summation Parameter: 600

Figure 5. Default View of Customer Dashboard

This is the default view of the customer dashboard, without any job or city filters applied. Here, the national average and city average salary are of course equal, as we are not filtering by any city. We have clustered the cities by having high salaries, moderate salaries, or lower salaries on average. This can be seen by the map in the top left of the dashboard, where each dot represents a city and is colored according to its salary cluster.



National Average Salary $35,456

City Average Salary $44,357

Hiring Demand 79,181

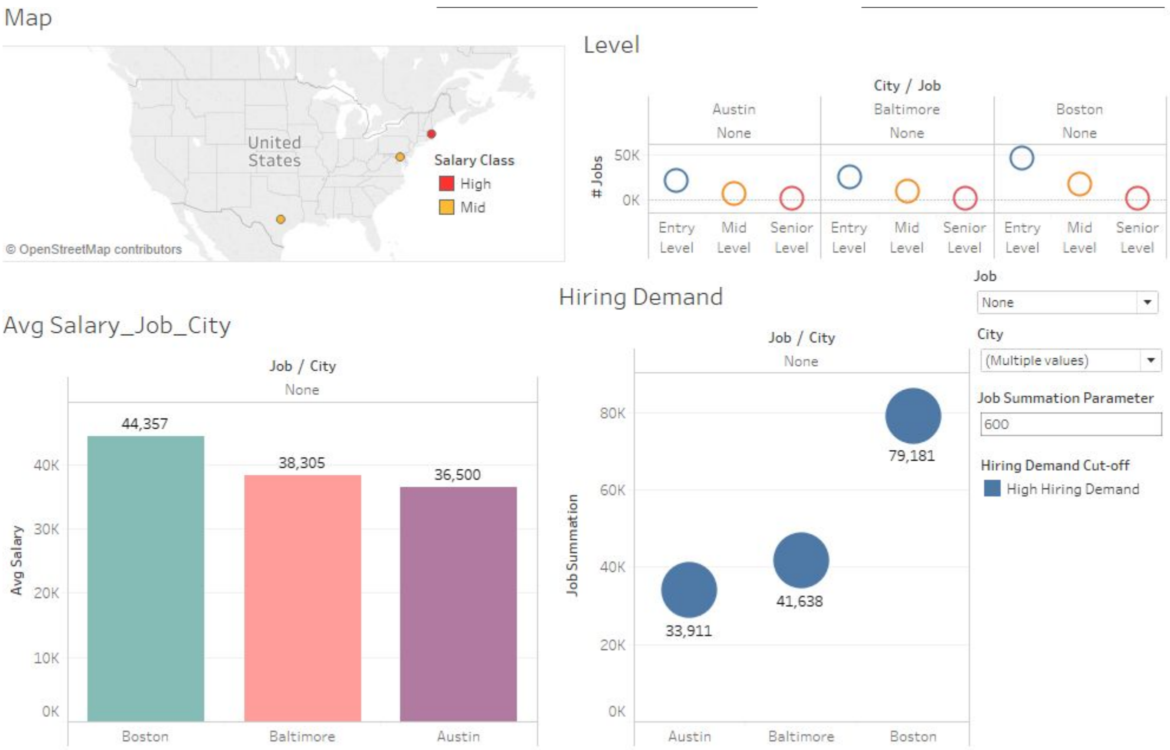
Job: None

City: Boston

Job Summation Parameter: 600

Figure 6. Customer Dashboard filtered by Boston

Here we have filtered by a single city: Boston. We can see that the average salary for this city is higher than the national average, and there are currently 79,000 job listings associated with the city. Things get more interesting as we add more cities.



Job: None

City: (Multiple values)

Job Summation Parameter: 600

Figure 7. Customer Dashboard filtered Three Cities

Here we have selected Boston, Baltimore, and Austin for comparison. We can see that Boston jobs offer higher salaries on average, as Boston is placed in the High salary class while Baltimore and Austin are in the Mid class.

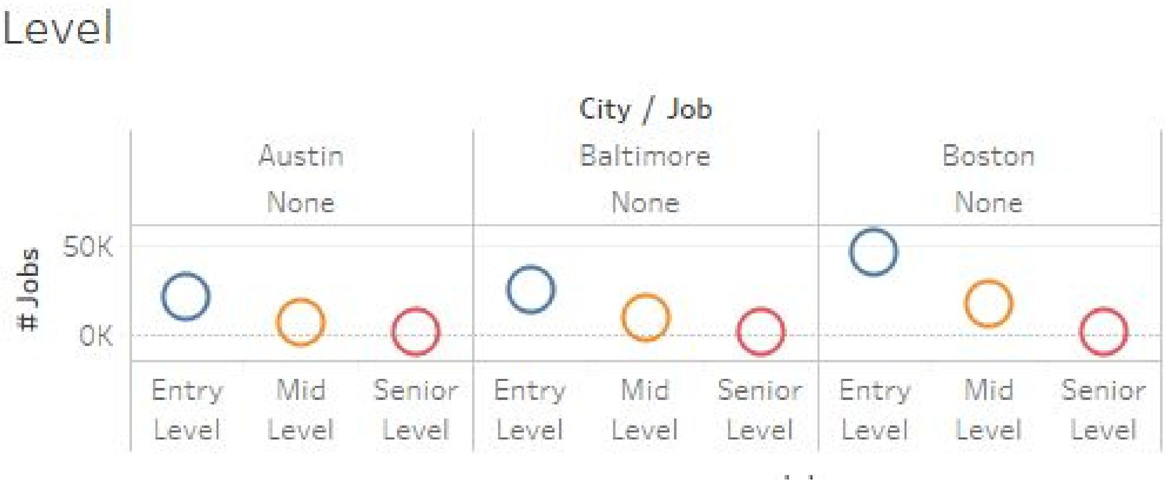


Figure 8. Demand Level of Three Cities

We see that for all three cities when no job filter is applied, low level jobs are in more demand than mid-level, and mid-level is in more demand than is high level.

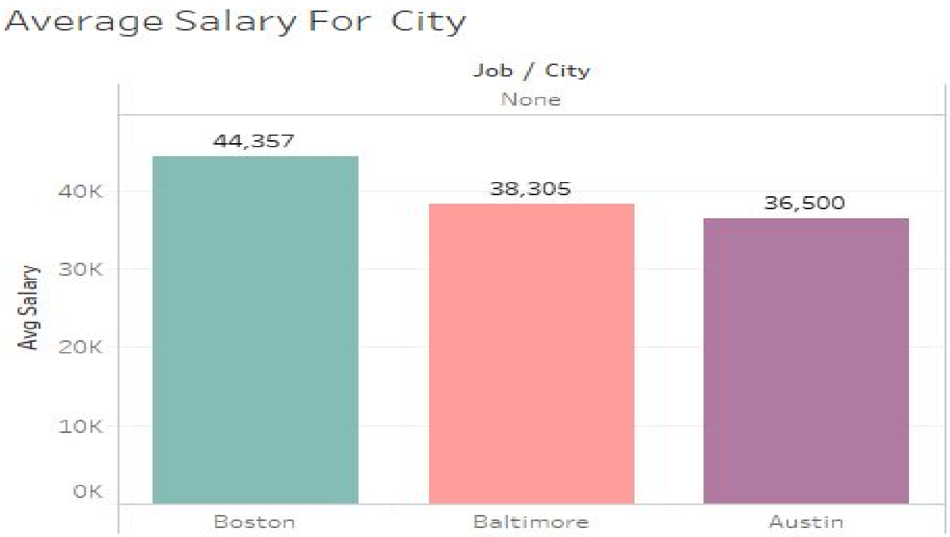


Figure 9. Average Salary of Three Cities

We can directly compare the average salaries of these three cities with a bar chart.

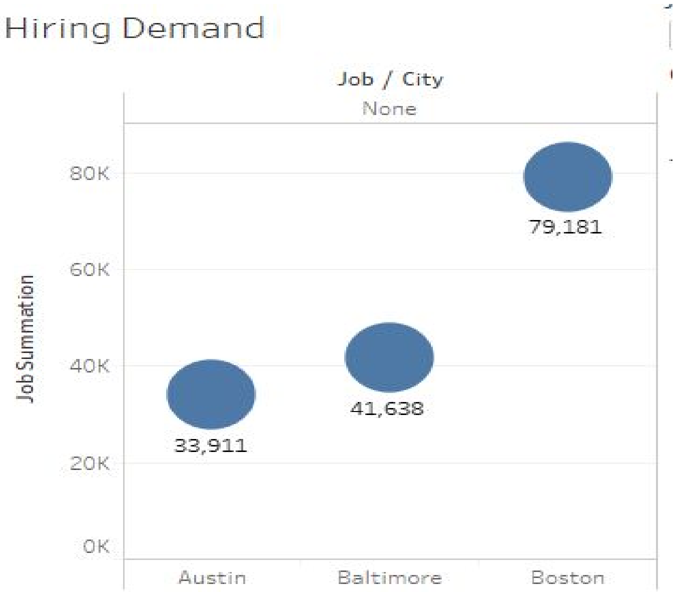


Figure 10. Hiring Demand in Three Cities

We can also see how the number of jobs at each city compares to the number of jobs offered at the others.

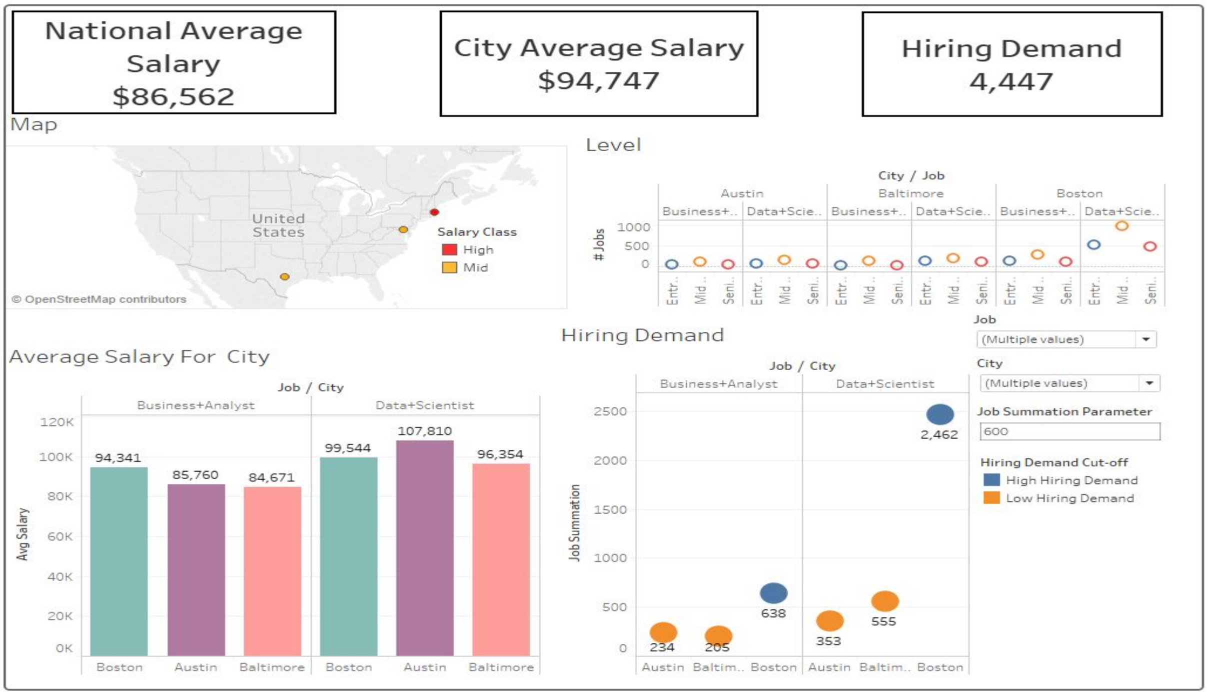


Figure 11. Customer Dashboard filtered by Job (BA& DS) and Cities

We can also compare these cities for multiple job titles. In this instance, we have chosen to compare Data Scientist and Business Analyst jobs at the three cities.

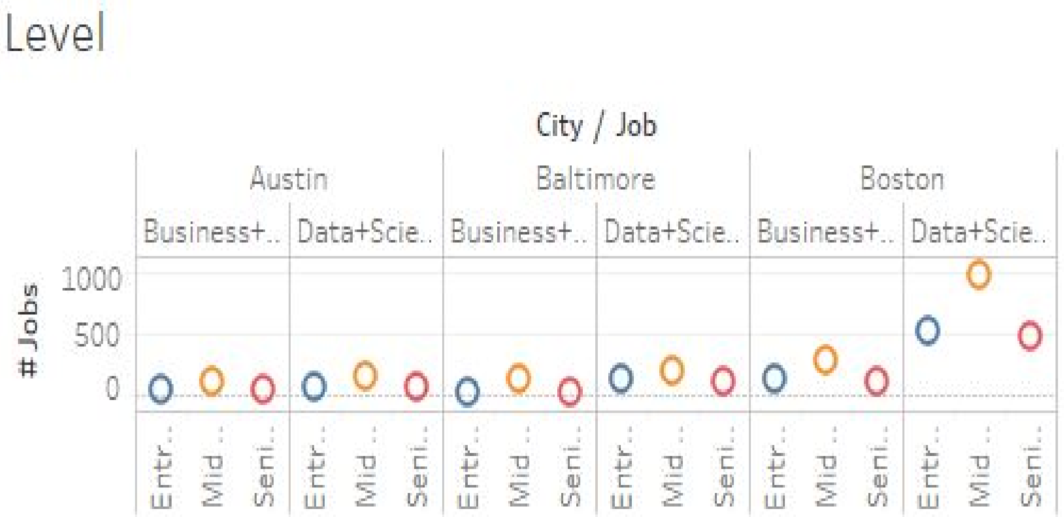


Figure 12. Level of BA & DS in Three Cities

Here we can see that for each city, mid-level jobs are in more demand for both Data Scientist and Business Analyst than are entry and high level jobs.

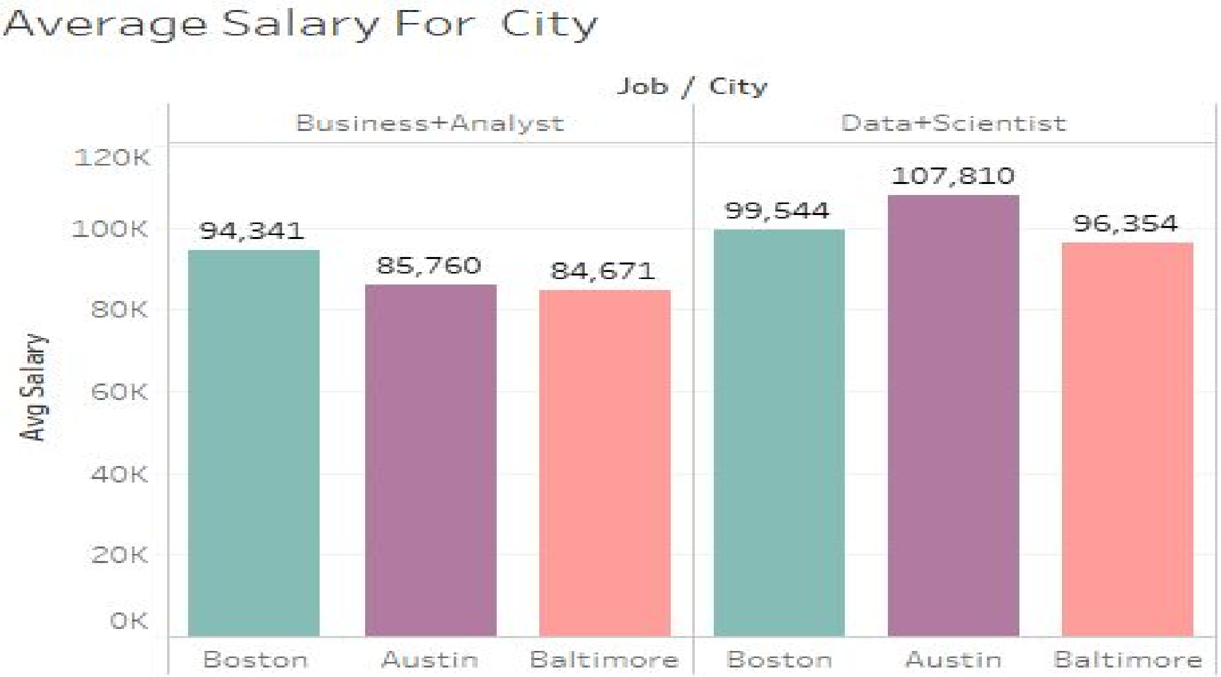
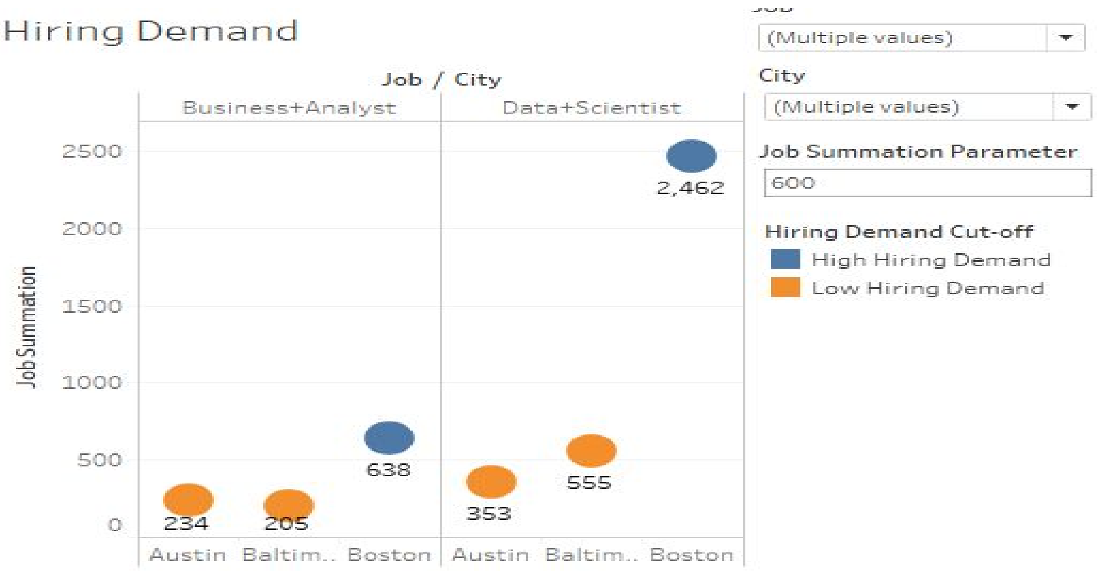


Figure 13. Average Salary of BA & DS in Three Cities

In figure 13, we can compare the average salary of business analyst and data science jobs for each city. We see that salaries for Data Scientists are higher than those of Business Analysts. Additionally, Austin has the highest salary for data scientists out of the three cities we've selected, and Boston has the highest salary for Business Analysts.



Job
(Multiple values)

City
(Multiple values)

Job Summation Parameter
600

Figure 14. Hiring Demand of BA & DS in Three Cities

We can also investigate the demand for these job titles across our three cities. We can set a cutoff for the number of jobs we consider to constitute a high demand. With a cutoff of 600 jobs selected, we see that Boston has a high demand for both data scientists and business analysts, while Austin and Baltimore have a low demand.

4.2. The Operational Dashboard

This dashboard, which will primarily provide benefits at the operational level, provides a summary of Jira commits associated with each active project at Indeed and allows project managers to monitor the progress of their team and the state of their project. This is done by providing project managers with KPIs such as Percent of Open Jobs and Change in Commit Rate, as well as allowing the project manager to quickly access the breakdown of what kind of commits are being made to their project and pick out which team members providing the most commits.

This dashboard provides easy access to the KPIs we have identified: The total commits made to a project, the change in commit rate, and the number of jobs left open. The total number of commits allows the manager to see how active their project is, while the change in commit rate allows them to see how progress on their project compares to activity last quarter. The percent of open jobs allows management to see the ratio of open jobs to completed jobs; a high open job percentage is a sign that there are many tasks that have not been completed and signifies that the project needs attention.

The Operational Dashboard: Default View

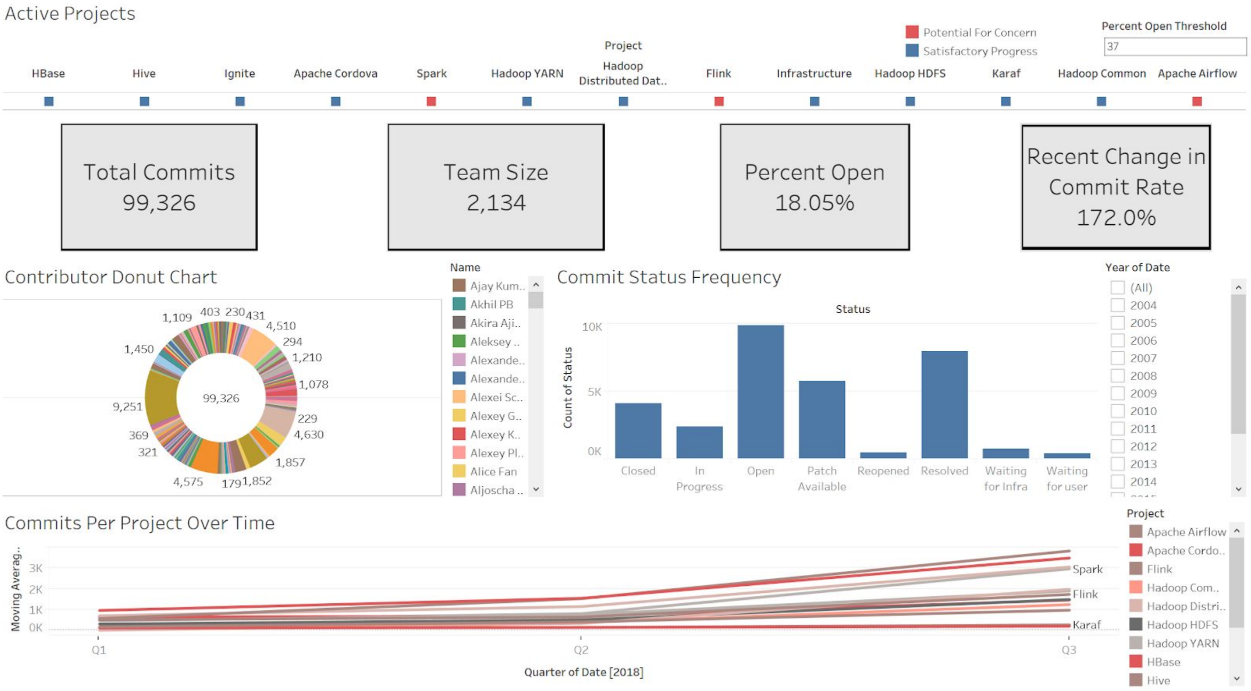


Figure 15. Default View of Operational Dashboard

Here we see the default view of the operational dashboard before any filters are applied. At the top of the dashboard is a list of all of Indeed's active projects. Below that we have the KPI's plainly visible in boxes. We then have a donut chart of the top contributors, next to which is a bar chart of the different types of commits being made. Finally, at the bottom of the dashboard we have a line chart of the moving average of the number of commits made over our specified time range, which by default is the current year.

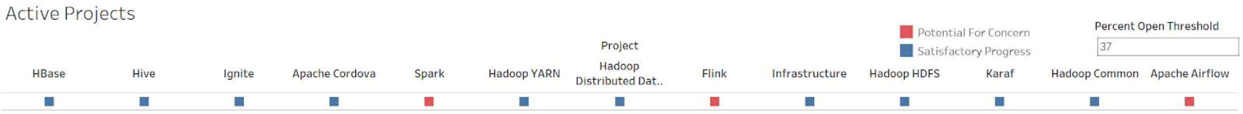
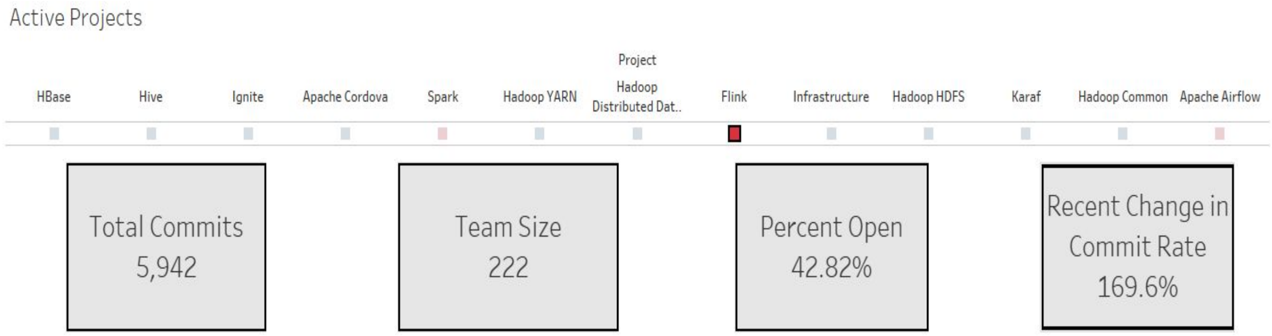


Figure 16. Active Projects

The active projects are color coded such that projects that need immediate attention are colored red. We define projects in need of attention as projects where over 40% of tasks associated with that task are open.

We will now demonstrate the value of this dashboard with a use case.

Let's suppose that the manager of the Flink project is concerned about the status of her project, as through the color coding she can immediately see that his project is in need of attention. In order to drill down and see details about her specific project, she can click on the icon below "Flink" the filter by her project:



Active Projects

Project

HBase | Hive | Ignite | Apache Cordova | Spark | Hadoop YARN | Hadoop Distributed Dat.. | Flink | Infrastructure | Hadoop HDFS | Karaf | Hadoop Common | Apache Airflow

Total Commits
5,942

Team Size
222

Percent Open
42.82%

Recent Change in Commit Rate
169.6%

Figure 17. The KPIs of Project Flink

After clicking on the icon, the manager can now see the KPI's of her project. She can see that nearly 43% of the tasks associated with her project are still open, which was why her project was flagged as needing attention.

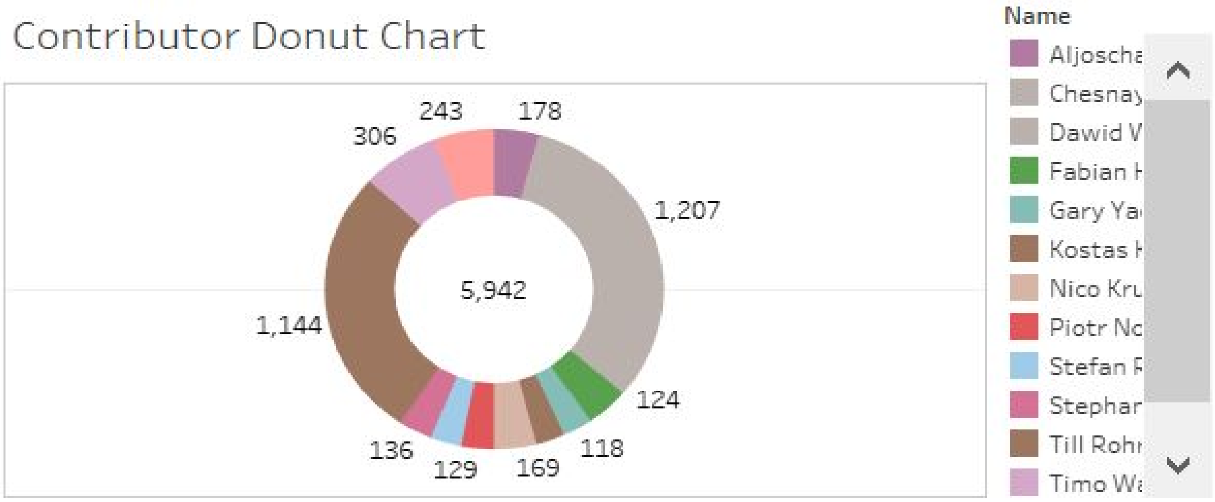
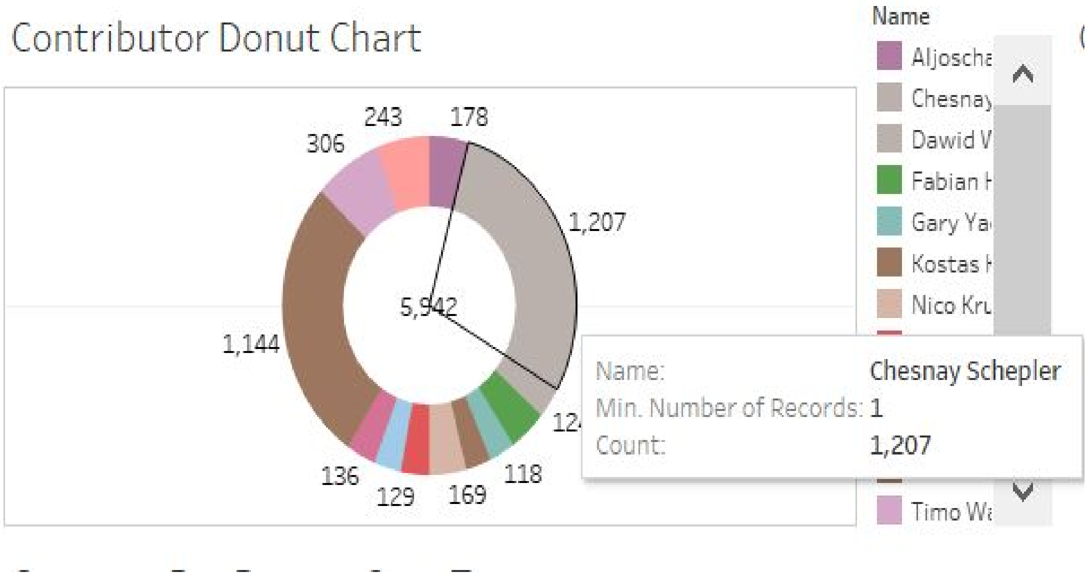


Figure 18. Contributor Donut Chart of Project Flink

Now that the manager has filtered by her project, she can also see a donut chart showing the proportion of commits made by the top contributors on her team. The team members are color coded, with a legend to the right of the plot.



Name: Chesnay Schepler
Min. Number of Records: 1
Count: 1,207

Figure 19. Chesnay Schepler's contribution in Flink

When she hovers her mouse over the largest portion of the chart, she sees that Chesnay is the team member who commits the most.

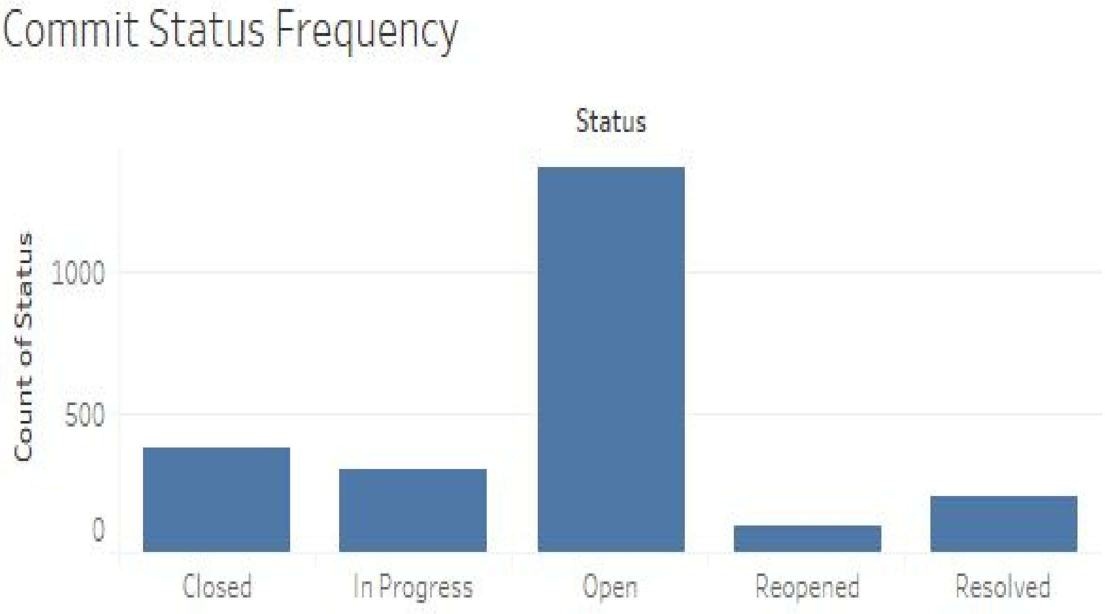


Figure 20. Commit Status Frequency

The manager is also presented with a bar chart showing the breakdown of types commits made by her team. As expected, most commits are opening new jobs.

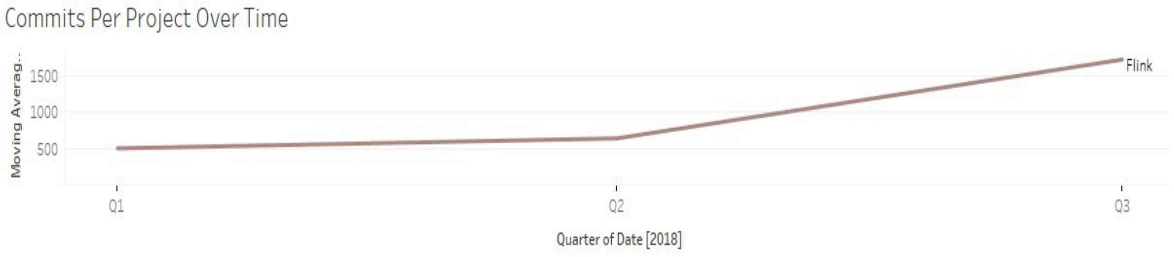


Figure 21. Commits Per Project Over Time

She can also see that while the number of commits rose only a small fraction between the first and second quarters of the year, but they have increased significantly over the last quarter.

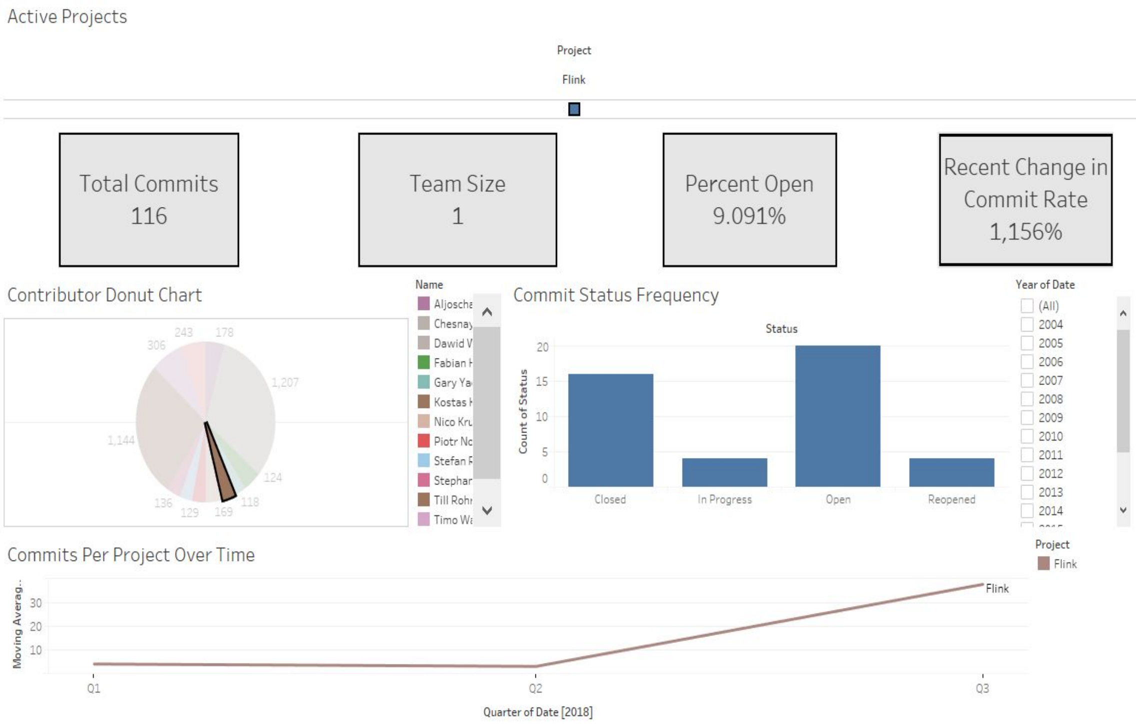
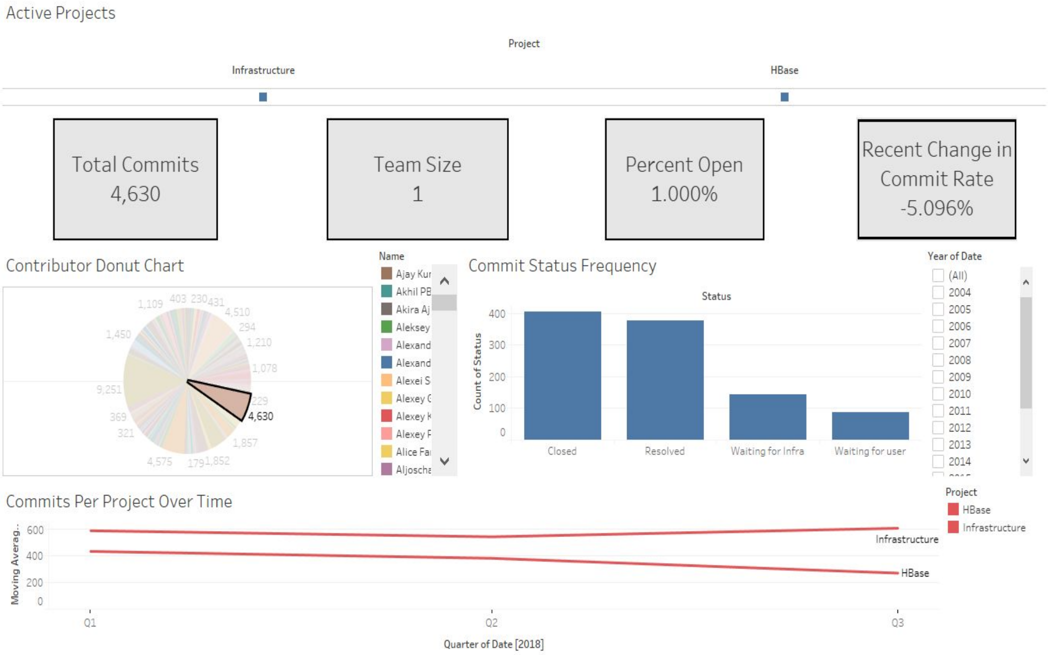


Figure 22. Operational Dashboard filtered by Kostas

So far, the manager has been investigating trends of her team as a whole. However, the dashboard also allows her to filter by individual team member. Here, she has selected Kostas. Now, she can see statistics about his behavior in specific. Kostas has only 9% of his tasks open, which is much lower than her team as a whole. She can also see that his commit rate has increased over 10 times over the last quarter, and while the majority of his commits are opening jobs, he has also closed a significant amount.



Total Commits 4,630

Team Size 1

Percent Open 1.000%

Recent Change in Commit Rate -5.096%

Figure 23. One Employee Has Multiple Projects On Hand

Finally, this dashboard also allows a manager to compare an employee's trend across different projects. Selecting Gavin, we can see that he is associated with two projects; HBase and Infrastructure. While his commit average has fallen for HBase over the last quarter, it has increased for Infrastructure.

4.3. Analytics

In order to demonstrate the benefits analytics can provide Indeed, we have also clustered Indeed's employees based on types of commits made and by number of projects worked on.

4.3.1. Cluster by Commit Type

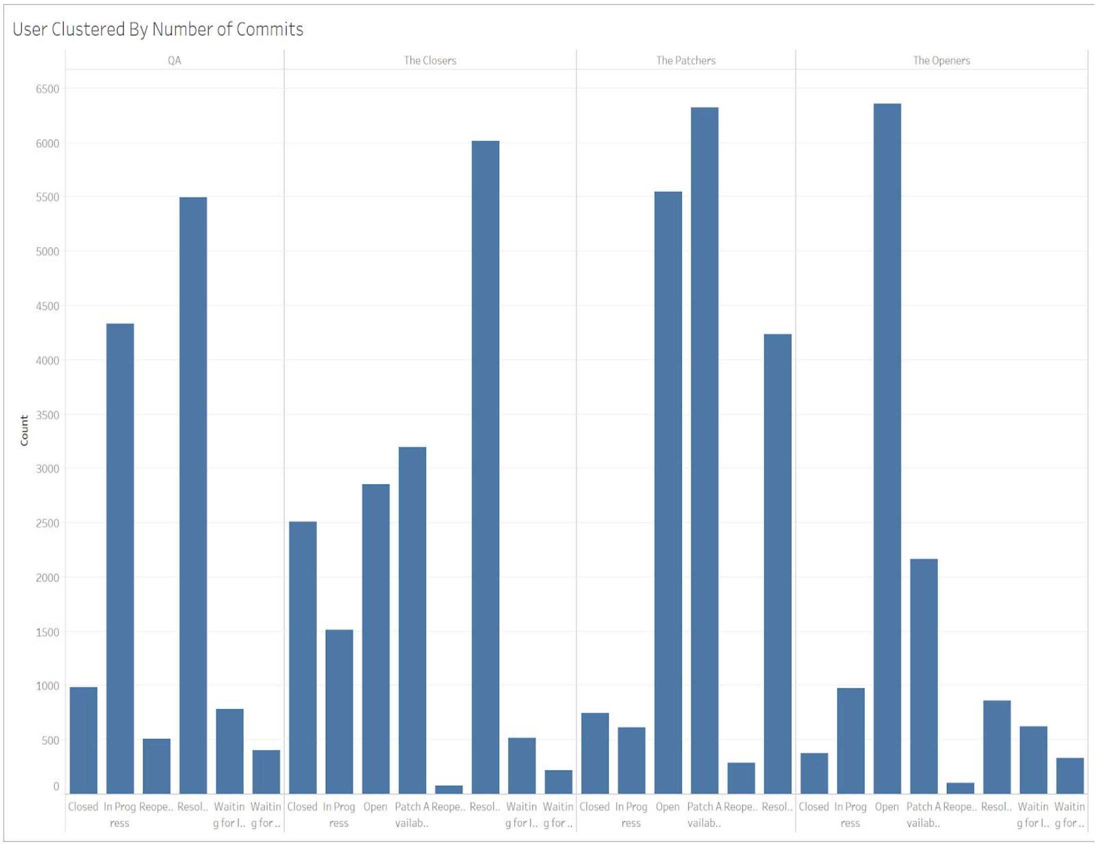


Figure 24. User Clustered By No. of Commits

When we cluster employees by commit type, we find four different types of commit behavior: 1) Employees who mainly open tasks but do little of any other type of commit, 2) those who open tasks, create patches, and resolve jobs, 3) those that mainly close jobs and resolve tasks, and 4) those who have In Progress and Resolved with most frequency.

Indeed regularly constructs teams with a relatively brief lifespan for specific purposes on an ad hoc basis. When Indeed puts together a team for a new project it is important to ensure that the team has the necessary skills and experience to result in a successful project. It is important have each of these clusters represented within the team; for example, the team needs people experienced with opening and assigning tasks, individuals experienced with putting out patches, and those who know how to close jobs. Our analytics solution allows employees to be automatically sorted into these clusters, so it is easy to make sure all of these skill types are represented when building a new team.

4.3.2. Cluster by Number of Projects

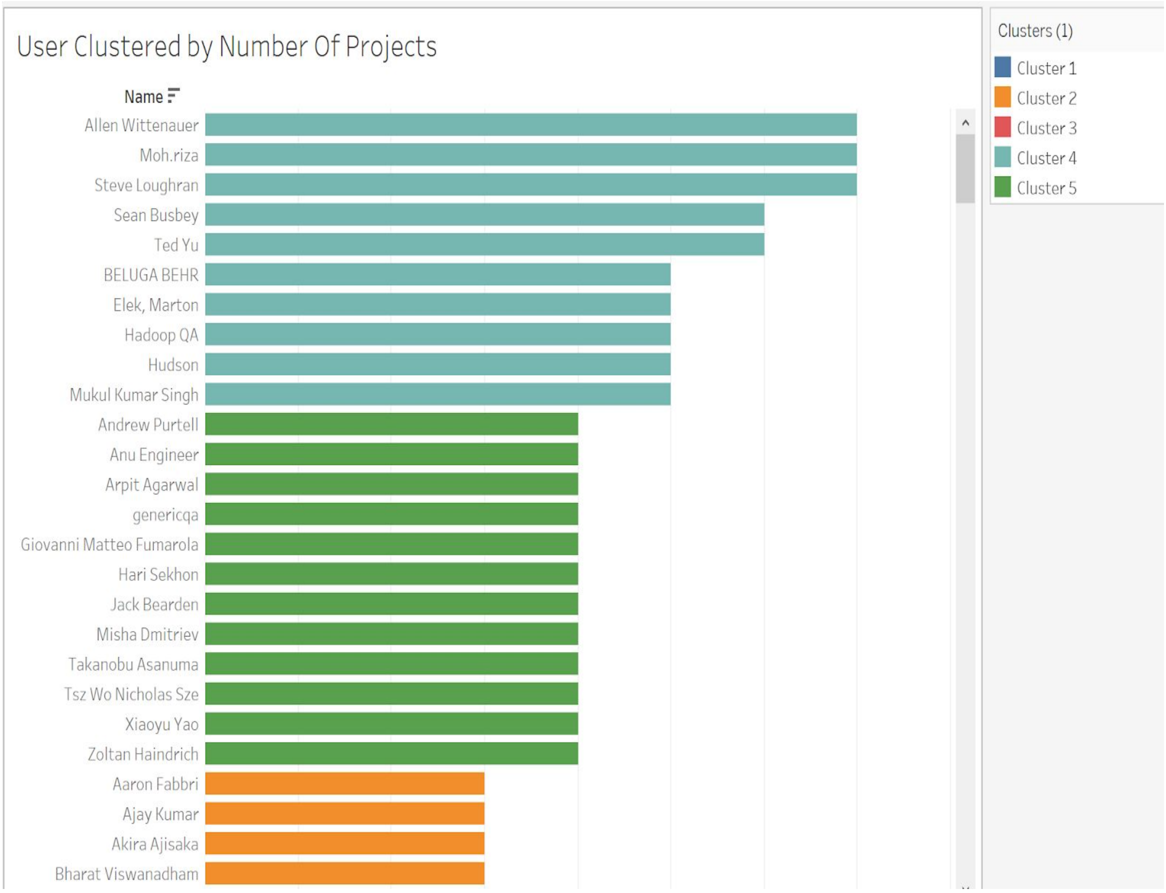


Figure 25. User Clustered By No. of Projects

We have also clustered employees by the number of projects they've worked on at Indeed, and discovered 5 broad clusters. When constructing a new team, in addition to the considerations discussed above, it is important to have at least a few team members that have worked on several projects in the past and are familiar with what it takes to produce a successful product for Indeed. However, upper management will not be intimately familiar with the history and experience level of all employees. By clustering employees in such a way, we provide management with a quick and easy way of accessing the experience of each employee.

5. BI implementation and recommendation

We have our dashboards ready, the problem now is how to successfully implement BI into the business. The most important thing in the process is business needs. It is important to have a business domain expert as a sponsor. On one hand, to make sure all what we are doing is solving business problems and making profit for Indeed company. On the other hand, to make sure we, as the BI team, deliver the right system for Indeed, users should be an important partner included the implementation process.

5.1. Managerial side

Kotter has defined 8 step process are as follows, and we will follow the steps to provide recommendations on the managerial side.

5.1.1. Create a sense of urgency

For preparing a change, a sense of urgency is a very powerful tool that helps people to win. Especially when the change is necessary for the organization and creates predictable value. For implementing our dashboard, Indeed needs to acknowledge the usefulness and the foreseeable future value of our new BI system. For example, Indeed's current customer dashboard comparing to LinkedIn, Handshake or other competitor companies, is not pretty enough, not ease of use and lack of competitive functionalities. When everyone notices the necessity of change, the sense of urgency can start building on itself.

5.1.2. Build a guiding coalition

After acknowledging the employees about the necessity of change. They can be influential and powerful in the process of implementing it. Also, they will know the business needs in Indeed company and are able to lead our dashboards in the right process to meet every business needs such as tracking productivity of employees and giving better customer experience. We meant business in Indeed to be improved from both internally and externally.

5.1.3. Form a strategic vision and initiatives

After making sure that our new BI system will be valuable to the company and the organization is craving the change, we recommend Indeed to think about a plan in this step. At the same time, start letting a small group of people to learn how use our new BI system in order to have a better understanding of our dashboards. Not only for them to feel the goodness and lead the implementation, but also for getting their opinions from a user's perspective which can be beneficial for future improvement.

5.1.4. Enlist a volunteer army

Our customer dashboard increases customer experience and gain loyalty and our employee helps to increase employees' working efficiency. A volunteer group who believes that they are implementing these values to business in Indeed here is the best drive of implementing the whole new BI system. As a result, a volunteer army must be brought in.

5.1.5. Enable action by removing barriers

Put in place the structure for implementation and keep checking for barriers. They do not use operational dashboard like ours to evaluate the productivity of employees before. Remove the current productivity evaluation method will provide a relatively positive environment for implementing change.

5.1.6. Generate short-term wins

Short-term wins provide the best motivation. One win always motivates the next one to happen. To generate short term wins, the simple way is to collect feedback from who has experienced our dashboards.

5.1.7. Sustain acceleration

There will be a dramatic increase of credibility after the first short-term wins which means there will be more opportunities to embrace visualization as a tool for enhancing efficiency and capturing them while the momentum persists is crucial.

5.1.8. Institute change

After all the process of implementing the new dashboards, it is very important to let everyone in Indeed start using them and finally get used to them. No matter employees or customers, this change in system should eventually replace their old habits of operating previous system.

5.2. Technical side

Data does not need to be perfect. However, in a company that mostly based on such massive amount of data like Indeed, we found that Indeed can improve their data quality in flowing ways.

5.2.1. Data integrity

Although missing a small part of data did not cause serious impact on our result, improving data integrity can still be beneficial. What is more, by improving data quality itself, BI teams' efficiency can be improved by having more time to spend on analysis instead of recovering missing data.

5.2.2. Data accuracy

Different from data integrity, data accuracy to us was a bigger problem. While missing data can be noticed and recovered in some level, inaccurate data can be invisible and not as noticeable missing ones among the whole dataset. Also, the missing data that we tried to recover has the possibility to be inaccurate. Improving data accuracy can help removing the threats that BI team are not able to realize sometimes not to mention to solve.

5.2.3. Data timeliness

In the employee table, we did not know if the data was a current database or just a piece of data warehouse. For an employee performance dashboard, dataset needs to be up-to-date. Employee dashboard is going to be used by managers, old data result can be confusing. Indeed has all information stored in their warehouse, but they need make sure they use the current datasets.

5.2.4. Data uniqueness

Exploring useful data that unique from what other companies will be beneficial for user experience. For example, Handshake provides information about if a company provides sponsorship for foreign employees. As an international student, I investigated job searching websites in a user's vision. I found their Handshake more satisfying only for highlighting companies that provide sponsorship and list them on the top of everything. Therefore, as a consultant for Indeed, we recommend Indeed to try to include more information that other companies are not currently including in their dataset or using in dashboard.

5.2.5. Data accessibility

By having access to more dataset, data analysts can do a better job. If we had more data on employee's other information like age, gender, race and marital status, we can correlate multiple factors and make a more reasonable assessment of employee performance. Also, improving data accessibility can help improving data integrity. If we had salary information for every level of a specific job, customer can consider it together with our current information which shows the number of positions opening and get more comprehensive information, therefore make better decision using our customer dashboard.

6. Conclusion

Our project aimed to provide BI solution for Indeed. We learned about Indeed’s current BI implementing status and compared with other companies’ BI in the same industry as well. We figured out how much Indeed as a company depends on its data, and on which stage Indeed is currently at in the BI maturity model. On its way of transforming into ‘Sage’, we defined the business problems are that they needed to be more competitive when facing customer and a clearer vision facing employees. We tried to find solutions by looking into BI framework, data source, data model and data visualization. After deeply analyzing current BI system in Indeed and exploring their available datasets, we put their datasets to better use and developed two dashboards accordingly.

As a brief summary of our BI product for Indeed, it includes one customer dashboard and an operational dashboard. In the customer dashboard, we included information like average salary and opening position for every job in every city which was not included in their current dashboard on indeed.com. In the operational dashboard, we enabled project managers to track the most current project performance which previously was not clear, as easy to use or as up to date while using their current BI product.

We highly recommend Indeed to use our dashboard as soon as possible for our dashboards really can improve business in Indeed. Speaking of the advantages of using dashboards, we recommend Indeed to keep an eye on their data quality as well to provide better resources for future upcoming analysis and dashboards. Moreover, Indeed needs to have a more business centralized HR structure for them to dig deeper into their clients and analyze them. The goal is to find out a more accurate business needs so that they can implement BI systems that are more comprehensive, more efficient and more useful in the future.

References

1. (2018) "About Indeed". Retrieved from https://www.indeed.com/about .

2. (n.d.) "Imhotep - Data Analytics Platform by Indeed". Retrieved from https://opensource.indeedeng.io/imhotep/

3. Indeed Engineering (2014, April 30). "Large Scale Interactive Analytics with Imhotep". Retrieved from https://engineering.indeedblog.com/talks/large-scale-interactive-analytics-with-imhotep/

4. Eckerson, W. (2011). *Performance dashboards measuring, monitoring, and managing your business* (2nd ed.). Hoboken, N.J.: Wiley.

5. (2018) "About Indeed Hiring Lab". Retrieved from https://www.hiringlab.org/about/.

6. The 8-Step Process for Leading Change - Kotter. (2018). Kotter. Retrieved from https://www.kotterinc.com/8-steps-process-for-leading-change/

7. Kotter's 8 Step process to successful change. (2016). Educational Business Articles. Retrieved from http://www.educational-business-articles.com/8-step-process/

8. Yeoh, William and Koronios, Andy 2010, Critical success factors for business intelligence systems, Journal of computer information systems, vol. 50, no. 3, Spring, pp. 23-32. Retrieved from https://pdfs.semanticscholar.org/7a66/7cdb124e404be1f0152260eade99b1f8d217.pdf

9. Rob, P. (2002). *Database system: Design, implementation, and management*.Cengage Learning. P. 592.

7. Appendix

Indeed Web Scraper

```
import urllib
import requests
import bs4
from bs4 import BeautifulSoup
import pandas as pd
import re
from string import ascii_uppercase

def URL_concat(job = '', city = 'Boston'):
    URL_first = "https://www.indeed.com/jobs?q="

    URL_mid = "&l="

    return URL_first + job + URL_mid + city

#########
list_of_professions = []

for letter in ascii_uppercase:
    html_professions =
requests.get('https://www.indeed.com/find-jobs.jsp?title='+str(letter
))
    soup_professions = BeautifulSoup(html_professions.content,
'html.parser', from_encoding="utf-8")

    for each in soup_professions.find_all('table', {'id':
'letters'}):
        for profession in each.find_all(class_='job'):
            list_of_professions.append(profession.text)

list_of_professions = [w.replace(' ', '+') for w in
list_of_professions]
```

```
list_of_professions.append('Data+Scientist')
list_of_professions.append('')
###############

cities = set(['New+York', 'Chicago', 'San+Francisco', 'Austin',
'Seattle',
    'Los+Angeles', 'Philadelphia', 'Atlanta', 'Dallas', 'Pittsburgh',
    'Portland', 'Phoenix', 'Denver', 'Houston', 'Miami',
    'Charlottesville', 'Richmond', 'Baltimore', 'Harrisonburg',
'San+Antonio', 'San+Diego', 'San+Jose'
    'Austin', 'Jacksonville', 'Indianapolis', 'Columbus',
'Fort+Worth', 'Charlotte', 'Detroit', 'El+Paso',
    'Memphis', 'Boston', 'Nashville', 'Louisville', 'Milwaukee',
'Las+Vegas', 'Albuquerque', 'Tucson',
    'Fresno', 'Sacramento', 'Long+Beach', 'Mesa', 'Virginia+Beach',
'Norfolk', 'Atlanta', 'Colorado+Springs',
    'Raleigh', 'Omaha', 'Oakland', 'Tulsa', 'Minneapolis',
'Cleveland', 'Wichita', 'Arlington', 'New+Orleans',
    'Bakersfield', 'Tampa', 'Honolulu', 'Anaheim', 'Aurora',
'Santa+Ana', 'Riverside', 'Corpus+Christi', 'Pittsburgh',
    'Lexington', 'Anchorage', 'Cincinnati', 'Baton+Rouge',
'Chesapeake', 'Alexandria', 'Fairfax', 'Herndon',
    'Reston', 'Roanoke'])

#################



def create_df(job=[''], cit = ['Boston', 'Houston']):

    out_df = pd.DataFrame(columns=["City", "Job",
                                   "Full-time", 'Part-time',
'Contract', 'Commission', 'Temporary', 'Internship',
                                  'Entry Level', 'Mid Level', 'Senior
Level', 'Top Employer_1', 'Top Employer_2',
                                   'Top Employer_3','Top
Employer_4','Top Employer_5', 'avg_salary'])
```

```
    ci = 0
    for city in cit:
        print(city)
        for j in job:
            ci += 1
            if j == '':
                job_temp = 'None'
            else:
                job_temp = j
            positions = {'Full-time':None, 'Part-time':None,
'Contract':None, 'Commission':None, 'Temporary':None,
'Internship':None}
            experience_levels = {'Entry Level': None, 'Mid Level':
None, 'Senior Level': None}
            top_employers = {'Top Employer_1': None, 'Top
Employer_2': None, 'Top Employer_3': None, 'Top Employer_4': None,
'Top Employer_5': None}

            #print(city)
            html = requests.get(URL_concat(j, city))
            soup = BeautifulSoup(html.content, 'html.parser',
from_encoding="utf-8")

            for each in soup.find_all('div', {'id': 'rb_Job Type'}):
                try:
                    title = each.find(class_='rbList').text
                except:
                    title = 'None'

            for each in soup.find_all('div', {'id': 'rb_Experience
Level'}):
                try:
                    exp = each.find(class_='rbList').text
                except:
                    exp = 'None'
            #for each in soup.find_all('div', id_= "rb_Job Type" ):
            for each in soup.find_all('div', {'id': 'rb_Company'}):
                try:
                    emps = 
each.find(class_='rbList').text#.replace('\n', '')
                except:
                    emps = 'None'

            for each in soup.find_all('div', {'id': 'SALARY_rbo'}):
                try:
                    sals = 
each.find(class_='rbList').text#.replace('\n', '')
                except:
                    sals = 'None'

            split_sals = sals.split('\n')
            split_emps = emps.split('\n')
            split_title = title.split('\n')
            split_exp = exp.split('\n')

            for k in positions.keys():
                for t in split_title:
                    if k in t:
                        positions[k] = int(t.split('
')[1].replace('(', '').replace(')', ''))
            for e in experience_levels.keys():
                for s in split_exp:
                    if e in s:
                        experience_levels[e] = int(s.split('
')[2].replace('(', '').replace(')', ''))
            for indx, emp in enumerate(top_employers.keys()):
                if (indx+1)*2 < len(split_emps):
                    top_employers[emp] = 
split_emps[(indx+1)*2].split(' (')[0]

            list_of_sals = []
            list_of_nums = []
            for s in split_sals:
                if '$' in s:
                    sal_and_num = s.split(' ')
                    sal = int(sal_and_num[0].strip('$').replace(',',
''))
                    num = int(sal_and_num[1].replace('(',
'').replace(')', ''))
                    list_of_sals.append(sal)
                    list_of_nums.append(num)

            list_of_nums_cut = []

            for indx in range(len(list_of_nums) -1):
                list_of_nums_cut.append(list_of_nums[indx] -
list_of_nums[indx+1])

            list_of_nums_cut.append(list_of_nums[-1])


            list_sal_times_num = [a*b for a,b in zip(list_of_sals,
list_of_nums_cut)]

            avg_sal = sum(list_sal_times_num)/sum(list_of_nums_cut)

            out_df = out_df.append(pd.Series([city, job_temp] +
list(positions.values()) + list(experience_levels.values()) +
list(top_employers.values()) + [avg_sal],
index=list(out_df.columns)), ignore_index=True)
        if ci % 100 == 0:
            print(i)
    return out_df

out_df = create_df(list_of_professions, cit=cities)

out_df.to_csv('out_df.csv')
```