1. Title -- Name of your website/application (Be creative!)

2. Introduction -- Explanation of project goals/target problem, description of application functionality, list of group members

**1. Motivation for the idea/description of the problem the application solves**

The success of a business venture is a reason for founders and investors to feel proud. Both entrepreneurs and investors are constantly looking for resources, strategies, and guidance that will put them ahead of their rivals. It is questionable whether having certain inherent skills is a prerequisite for success as an entrepreneur or if those skills can be learned (e.g. through formal business education). Innovative teams and leaders in market research rely on Crunchbase’s private company data to build powerful internal databases and research insights in respective industries. Furthermore, it is highly challenging to gauge the importance of external factors like the sector a firm competes in, the region where its headquarters are, or the amount of rivalry in a specific sector and its sub sectors.

Even though financing investments are one of the most frequently discussed topics in the world of investing and business, there aren't many studies that show how businesses may encourage funding investments from investors. Knowing what investors are searching for, or the elements that influence investing behavior, can help organizations better understand how to boost their prospects of securing financial investments from investors. To comprehend investment behavior in general, many research have been conducted. Some theories on what motivates investment include psychological and geographic characteristics, funding investment experiences, and even heredity. This project gives a broad overview of the current and future uses of the Crunchbase database as an original data source on cutting-edge businesses and start-ups around the world and assists users in identifying innovative organizations through dynamic dataset comparisons, tools, and insightful suggestions such as :

## **Access Live Company Data** : As our application develops, observe and detect market patterns. Our worldwide coverage of businesses at all stages, from pre-seed to late-stage, allows you to see how much money companies in a particular field are raising and who are the top players in specific industries. Receive suggestions and updates delivered to you regularly depending upon your activity.

## **Data in a Flexible Format :** You can use our application to find information about a list of businesses you're researching. Detailed industry-specific landscape research that includes real-time data on investment rounds, acquisitions, geography, and industry.

**2. List of group members, email addresses, and GitHub usernames**

Joseph Poirier - [jcpoir@seas.upenn.edu](mailto:jcpoir@seas.upenn.edu), jcpoir

Lisa Friedmann - [lisafrie@seas.upenn.edu](mailto:lisafrie@seas.upenn.edu), lisafrie

Purvansh Jain - [purvansh@seas.upenn.edu](mailto:purvansh@seas.upenn.edu), purvansh-jain

3. Architecture -- List of technologies, description of system architecture/application

**8. List of technologies you will use**

* Colab - Use Pandas/SQL for EDA, pre-processing and creation of data according to schemas
* React - Frontend (client side)
* [React-simple-maps](https://www.react-simple-maps.io/) - A declarative API for generating map-based visuals
* [Chart.js](https://www.chartjs.org) - A flexible and interactive charting library
* Node.js - Backend (server side)
* MySQL - Backend (Database querying)
* AWS - Hosting the database
* [Netlify](https://www.netlify.com/) - Deploying the website
* Github - Version control

**2. List of features you will definitely implement in the application**

* Graphs to describe data (e.g., line graph to show funding amounts over time, with drop-down to filter by industry, region, operational status; bar graph to show funding rounds for selected company)
* US Map highlighting one KPI, e.g. number of companies per state
* Filtered tables (search by name, filter by status, etc.) for company and investor
* Hyperlinks between pages

**3. List of features you might implement in the application, given enough time**

* Option to display different KPIs on map
* Drop-down to choose country for map
* Randomized fact of the day on homepage ( Related to successful entrepreneurs )
* Additional graphs to describe data (e.g., rating of investors based on amount of money spent relative to investor with highest investments, shown as progress bar or similar)

**4. List of pages the application will have and a 1-2 sentence description of each page**

* **Page 1 -** Homepage: quick overview of relevant summary statistics, such as number of companies, funding and acquisitions by year, with hyperlinks to other pages; statistics on funding categorized by industry type and current operational status depicted graphically (line chart over time)
* **Page 2 -** Dashboard: map displaying number of companies founded by US state. Potentially with option to show additional parameters, like number of financing rounds, number of acquisitions, dollar amount raised and / or option to display map of non-US based companies
* **Page 3 -** Companies: filter companies based on certain criteria, e.g. location, time of foundation, funding rounds, investors. Clicking on a state in the map on the dashboard page will bring users to this page and automatically apply the selected state as a filter. Clicking on a specific company will open a detailed view of the company below the filtered list, showcasing statistics about the selected company. It may have bar graphs about the different rounds of funding, and data specific to the company in question (location, founded date, industry type, acquisition, current status and parent company if any)
* **Page 4 -** Investors: similar to page 3, this page allows filtering of investors based on certain criteria, e.g. location, round type, investments made and shows a detailed view of one selected investor (potentially including relative status of investor based on money spent)

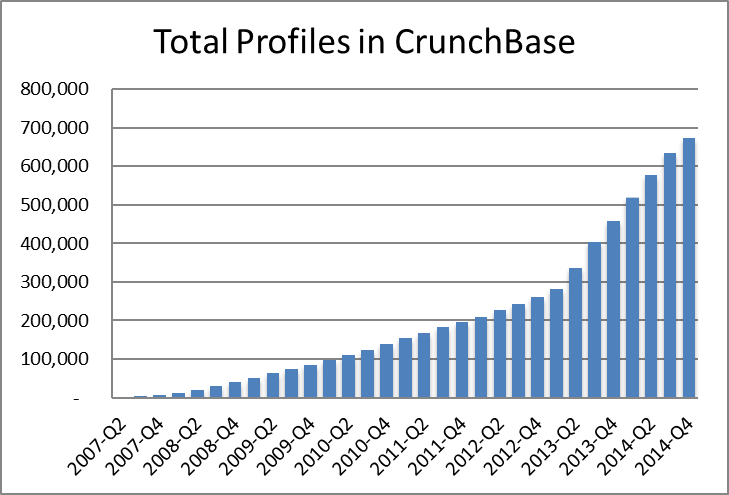
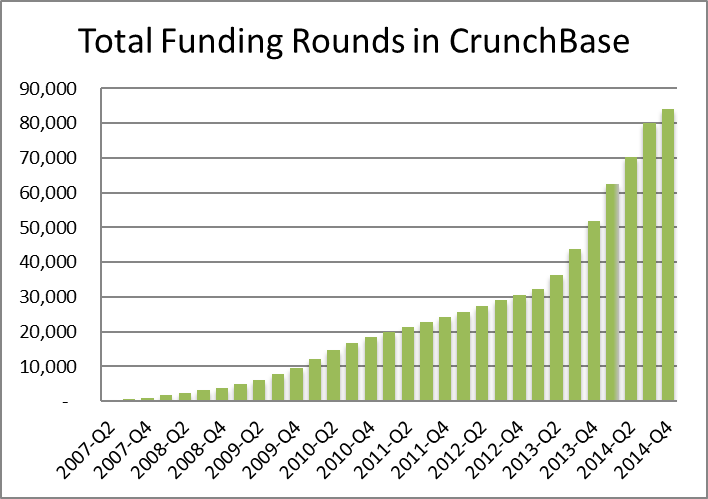
4. Data -- For each dataset, a link to the source, a description of the data, relevant summary statistics, and an explanation of how you use the data

1. **Description of the dataset:**

The dataset was found on [Tableau Sample Datasets](https://public.tableau.com/app/resources/sample-data?qt-overview_resources=1) and comes from the [Crunchbase](https://www.crunchbase.com/home) website.

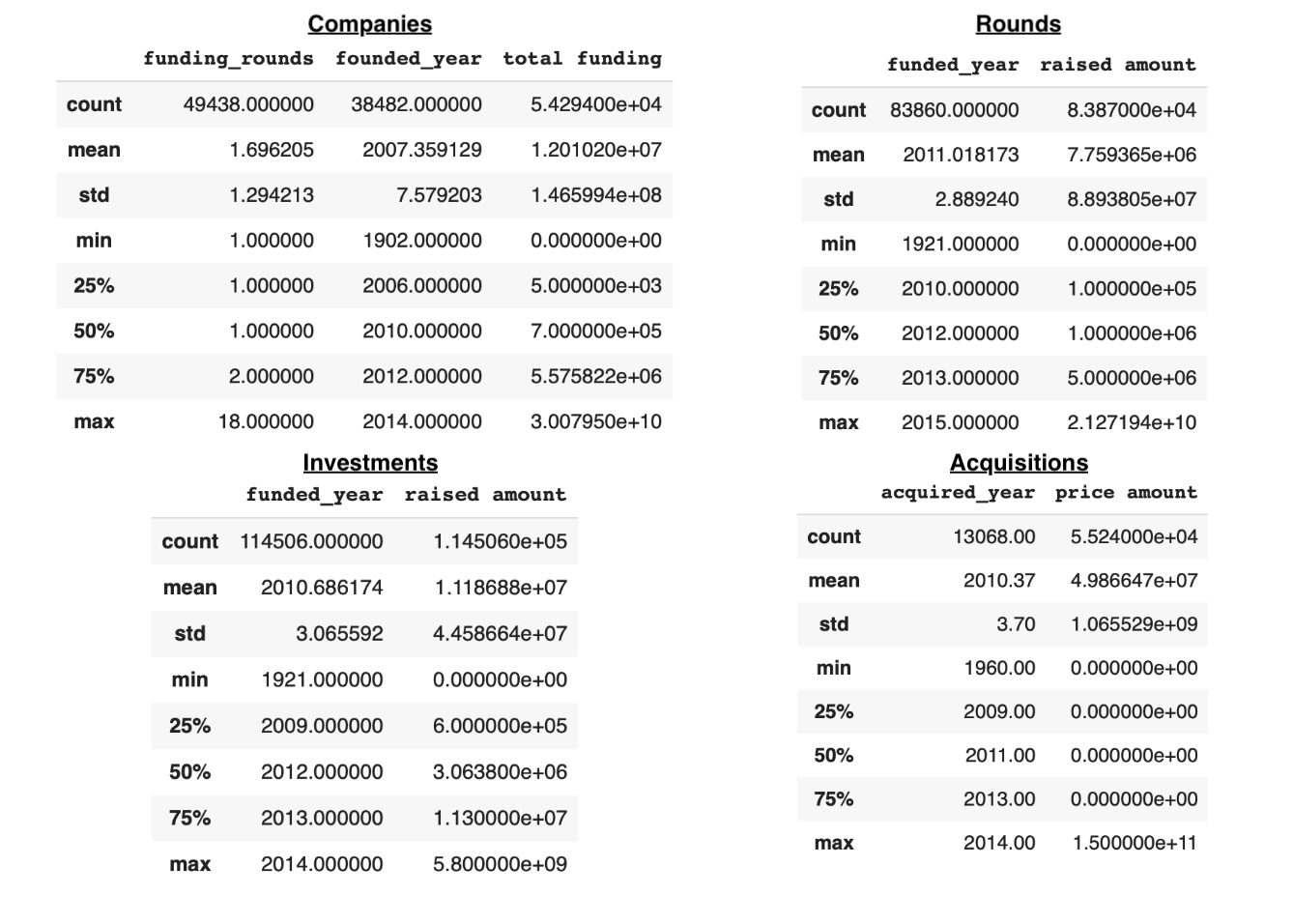
It contains information about companies with recorded fundraising 1960 and 2014. It has 4 separate tables:

* Companies: lists name, category, url, location, funding by company and other company characteristics
  + 49,439 rows, 18 columns
* Rounds: for each funding round for each company, lists type, date and amount
  + 83,871 rows, 16 columns
* Investments: lists information about the investors of each company, including name, location, round, date, amount invested
  + 114,507 rows, 24 columns
* Acquisitions: similar to Investments, this table lists information about acquisitions, including name of acquirer, location, date, price
  + 13,071 rows, 22 columns

Summary statistics:

Summary statistics of several attributes (e.g. report mean, standard deviation):

Calculated using Pandas ( for EDA)



**7. Explanation of how you will clean and pre-process the data**

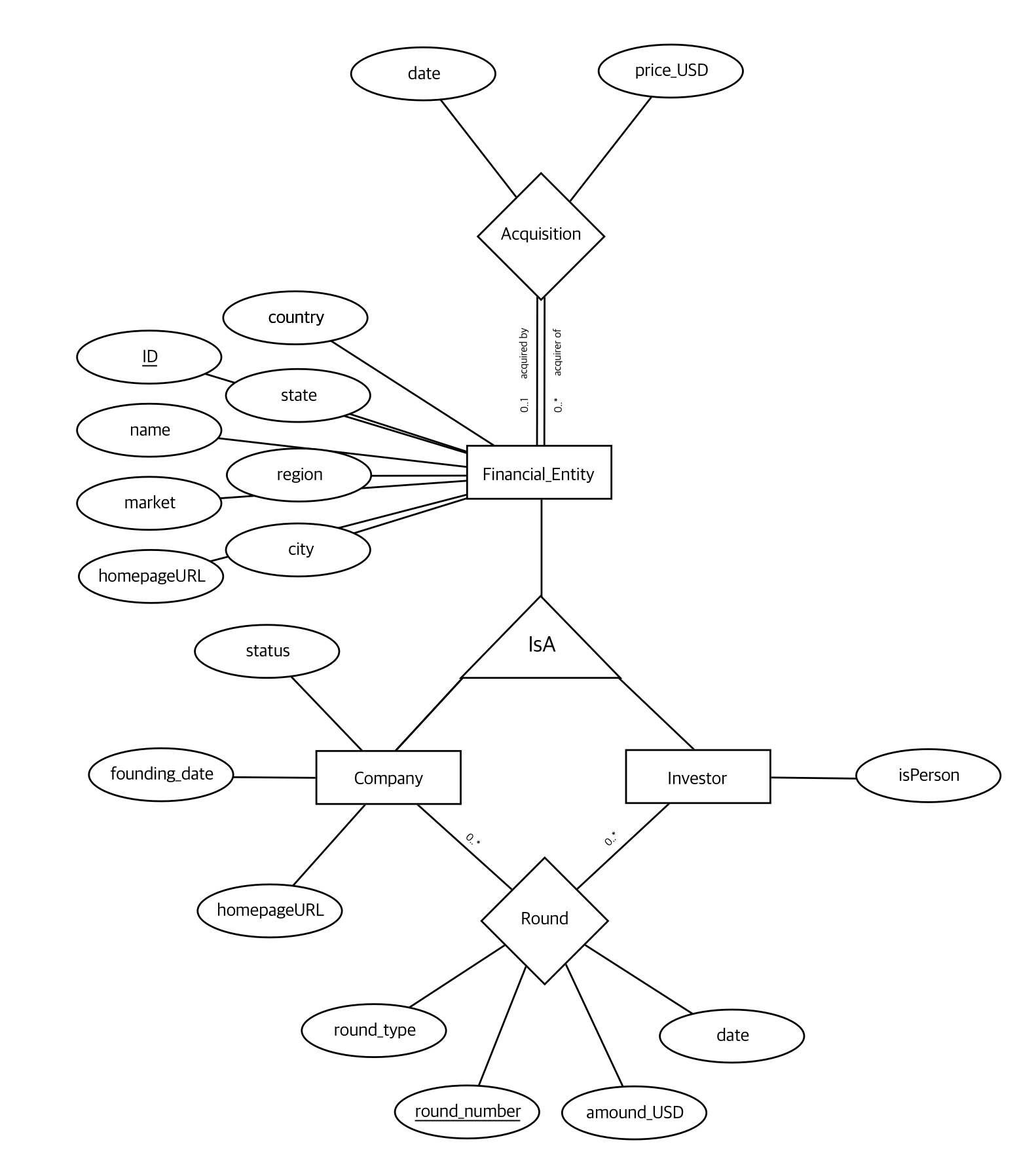
We initially begin with 4 datasets - Company, Investment, Acquisition and Round. After performing EDA on the data, we realized that some columns have missing or incorrect data, or are represented by incorrect data types. Furthermore, there is redundancy due to some columns being repeated between datasets and some rows are duplicated or have incorrect values. Based on these findings and our proposed schema, we will clean and pre-process the data as follows:

* **Company:** The original column permalink has two duplicate values, which will be resolved by aggregating the data into 1 row. Then this column will be renamed to CompanyID. Funding information (funding\_round, funded\_at, last\_funded\_at) will be dropped to eliminate redundancy (this all can be found in Round). Additionally, all calculated date columns (founded\_month, founded\_quarter, founded\_year) will be dropped, since they can be calculated from founded\_date. Some columns will be renamed to improve interpretability (for example, founded\_at will be changed to founding\_date). Company\_category\_list will be reformatted to be comma-separated and stripped to remove white spaces. Some founded\_at dates are incorrect, so they will be changed to nulls. Missing status values will be replaced by ‘Unknown’. Other missing values will be kept as nulls.
* **Investment:** This dataset has redundant information from Company and Round which will be dropped. The investor related columns will be renamed in accordance with the schema. investor\_category\_list will be reformatted to be comma-separated and stripped to remove white spaces. Investor\_permalink will be used as the identifying column here (similar to permalink in Company) and therefore we will remove duplicates. Additionally, the Investor\_permalink contains information about whether the row refers to a person or an organization. This information will be extracted from this column and be stored in a new column called isPerson. Missing values will be kept as nulls.
* **Acquisition:** Based on our EDA, we realized that an acquirer can be a Company, an Investor or exclusively in Acquisition. This dataset has redundant information from Company which will be dropped. The acquirer related columns will be renamed in accordance with the schema. acquirer\_category\_list will be reformatted to be comma-separated and stripped to remove white spaces. acquirer\_permalink will be used as the identifying column here (similar to permalink in Company) and therefore we will remove duplicates. Missing values will be kept as nulls.
* **Round:** This dataset has redundant information about Company and Investor, which we would drop. We will keep the company\_permalink and investor\_permalink as these along with a funding\_round\_ permalink will uniquely identify a round. Additionally, all calculated date columns (funded\_month, funded\_quarter, funded\_year) will be dropped, since they can be calculated from funded\_date. Some funded\_date values are incorrect, so they will be changed to nulls. Missing values will be kept as nulls.

Based on these datasets, we would then create Financial\_entity by merging Company, Investment and Acquisition and keeping the attributes common to all three. We would also remove all duplicates. The remaining subclass attributes can be extrapolated into their own tables in accordance with the schema.

5. Database -- Explanation of data ingestion procedure and entity resolution efforts, ER diagram, number of instances in each table, normal form and justification

**5. Relational schema as an ER diagram**



**6. SQL DDL for creating the database**

Financial\_Entity(ID int, name varchar NOT NULL, market varchar, country varchar, state varchar, city varchar, IsA enum(“company”, “investor”, “unk”))

Company(companyID int, founding\_date datetime, homepageURL varchar, status enum(“operating”, “closed”, “acquired”), FOREIGN KEY CompanyID REFERENCES Financial\_Entity(ID))

Investor(investorID int, isPerson boolean, FOREIGN KEY InvestorID REFERENCES Financial\_Entity(ID))

Round(companyID int, investorID int, round\_number int, round\_type varchar, amount\_USD int, date datetime, FOREIGN KEY companyID REFERENCES Company(companyID), FOREIGN KEY investorID REFERENCES Investor(investorID))

Acquisition(acquirerID int, acquiredID int, date datetime, price int, FOREIGN KEY acquirerID REFERENCES Financial\_Entity(ID), FOREIGN KEY acquiredID REFERENCES Financial\_Entity(ID)

6. Queries -- Examples of at least 5 queries in your application and explanations of how they’re used. Please report the queries you think are most complex. 7. Performance evaluation -- recorded timings before and after optimization(a table might be useful here), descriptions of events being timed, and explanations of why caching, indexing, and other optimizations improved timings. For example, record how long it takes to execute queries, respond to client requests, and execute any other common tasks.

8. Technical challenges -- List of technical challenges and how you overcame them