CC5212-1

Procesamiento Masivo de Datos

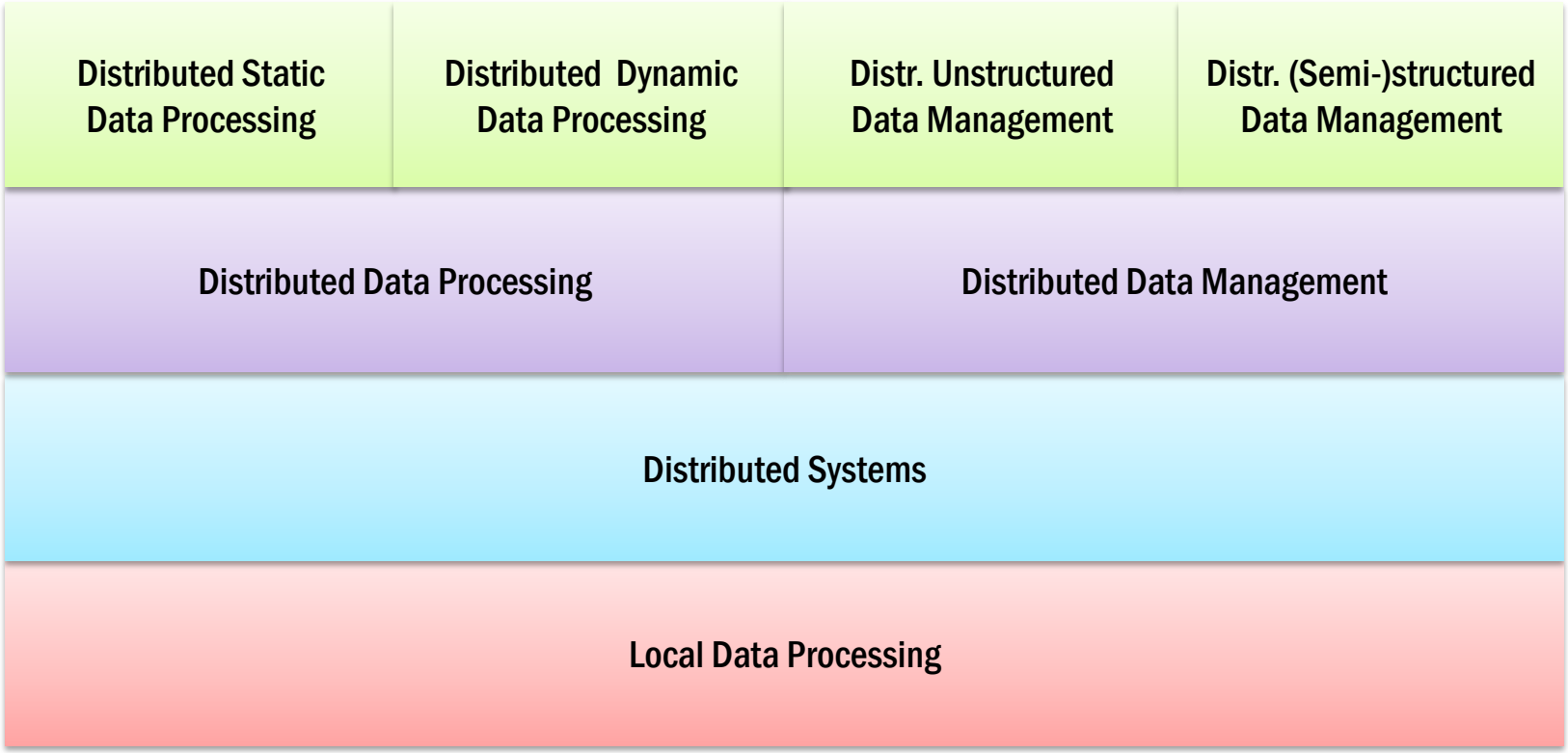
Otoño 2023

Lecture 12

Conclusion

Aidan Hogan

aidhog@gmail.com


Distributed Static Data Processing
Distributed Dynamic Data Processing
Distr. Unstructured Data Management
Distr. (Semi-)structured Data Management
Distributed Data Processing
Distributed Data Management
Distributed Systems
Local Data Processing

Distributed Systems

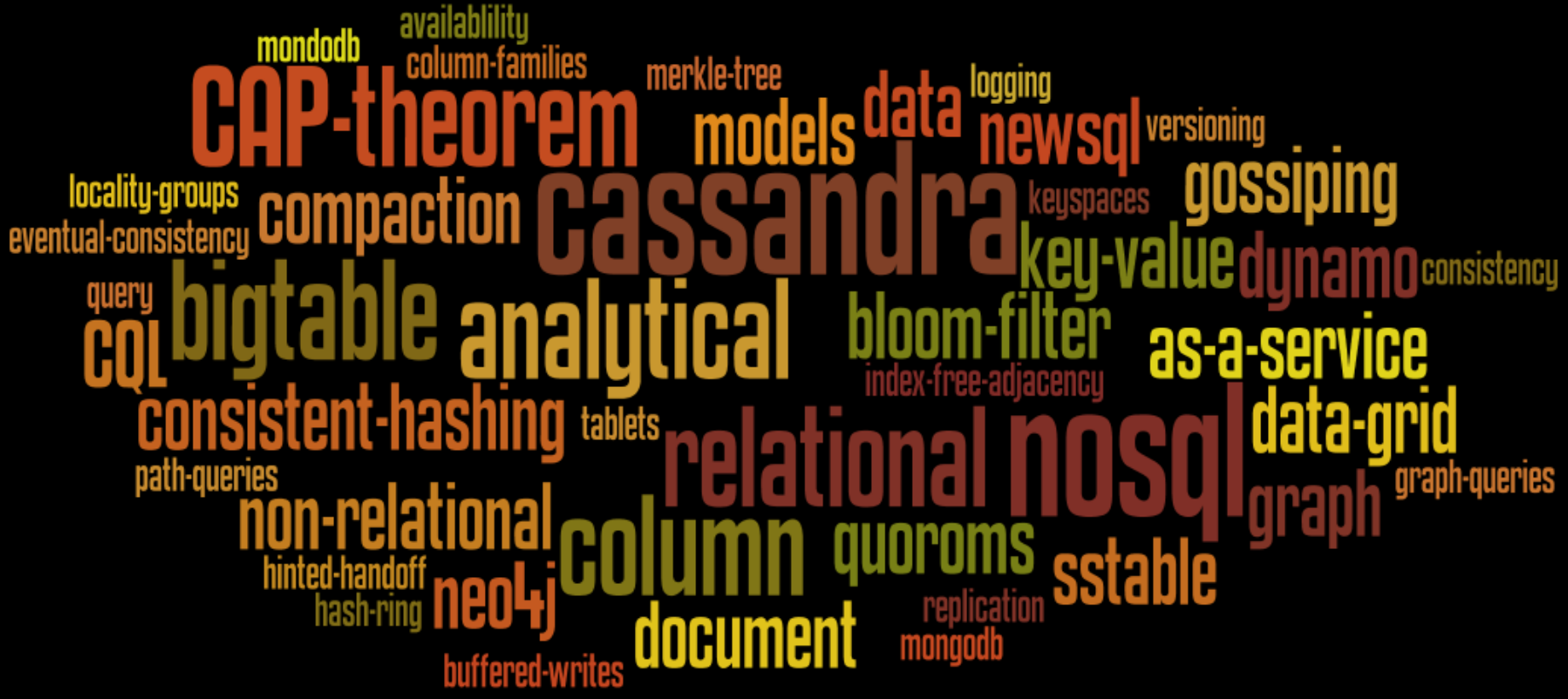
Hadoop/MapReduce/Pig/Spark:
Processing Un/Structured Information

Information Retrieval:

Storing Unstructured Information

NoSQL:

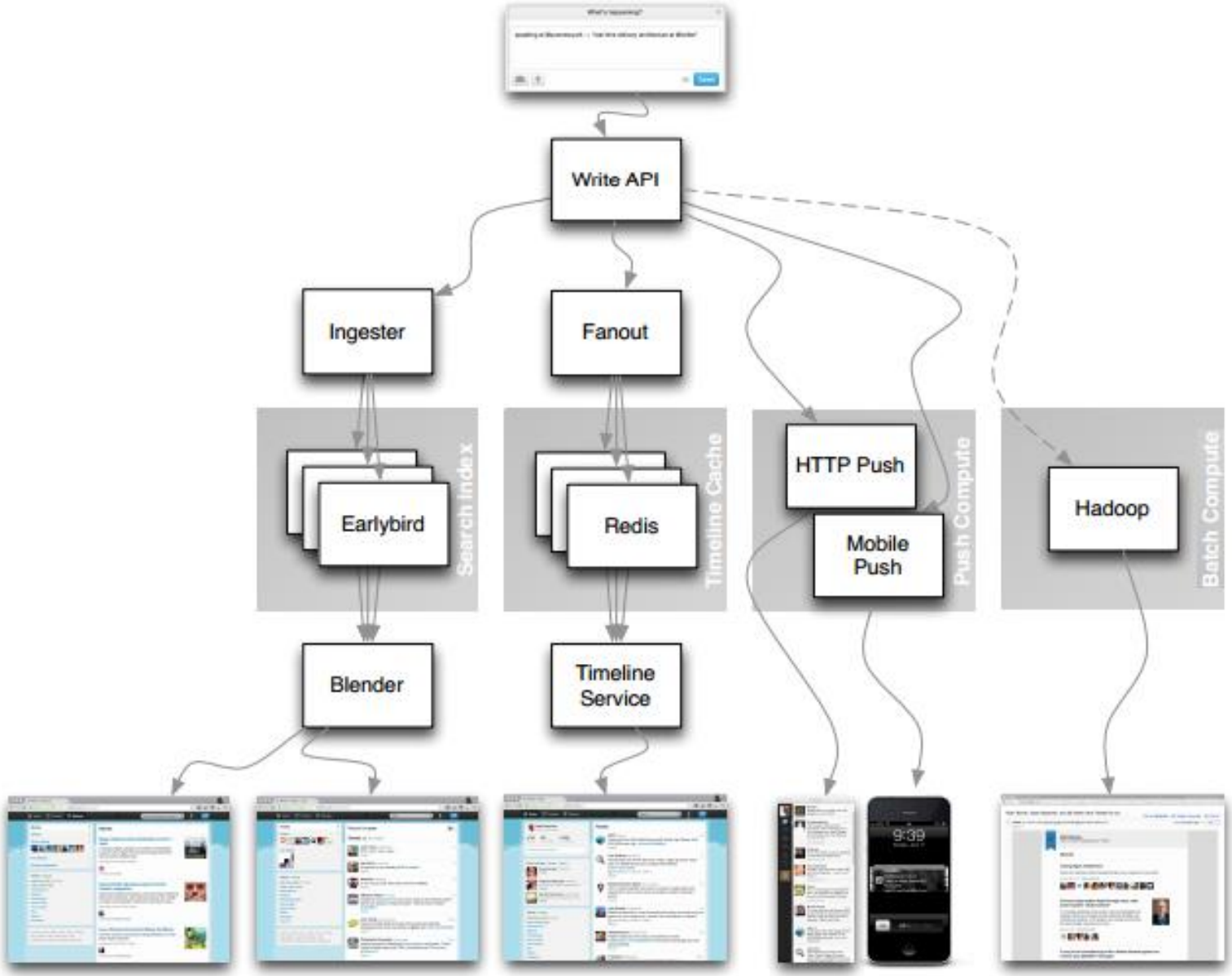
Storing (Semi-)Structured Information



Full-Circle

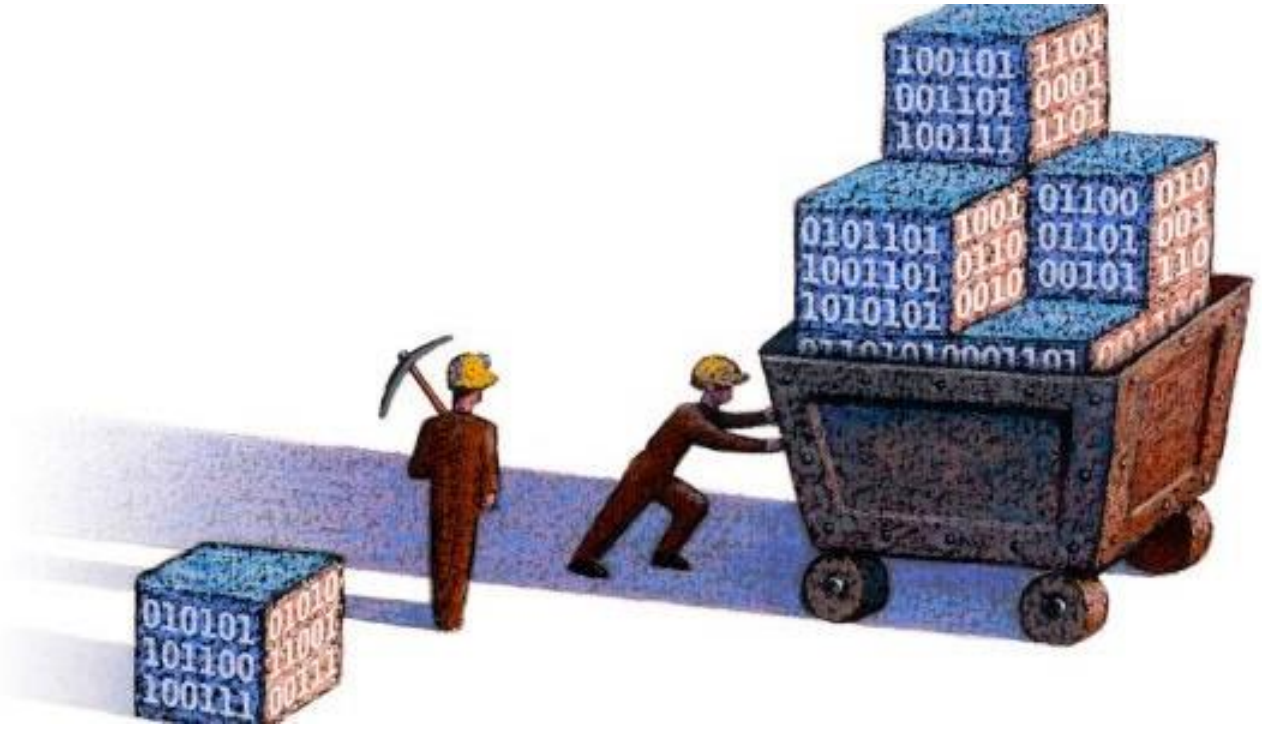
The value of data …

Twitter architecture

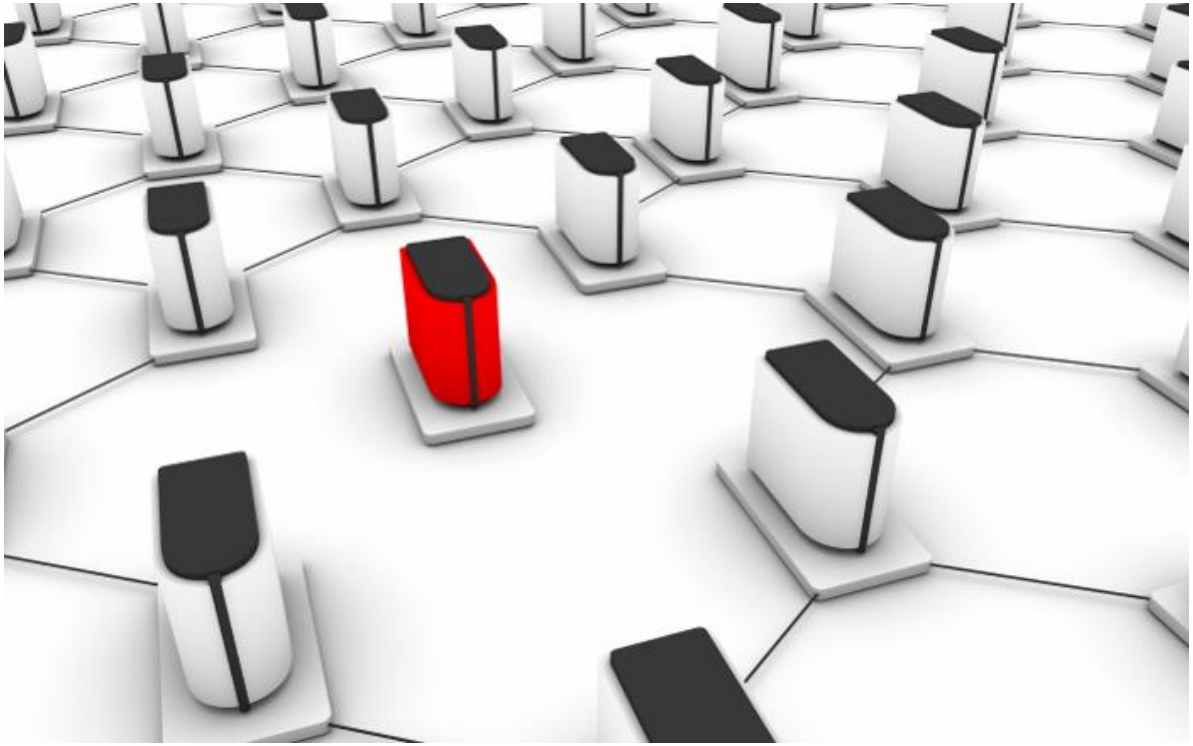


Generalise concepts to …

Working with large datasets

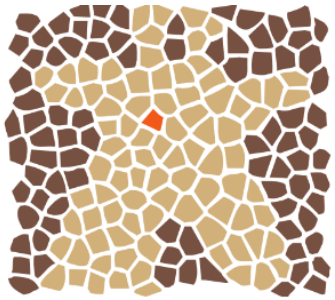


Value and challenges of distribution



Frameworks

- For Distrib. Processing



- For Distrib. Storage

"Data Science"

Harvard Business Review

DATA

Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

FROM THE OCTOBER 2012 ISSUE

“Data Scientist” Job Postings (2016)

Here are the top 10 in-demand skills for data scientists:

Skills	Job skill appears in	% of jobs with skill
SQL	1987	56%
Hadoop	1713	49%
Python	1367	39%
Java	1287	36%
R	1120	32%
Hive	1099	31%
Mapreduce	768	22%
NoSQL	657	18%
Pig	561	16%
SAS	560	16%

http://visit.crowdflower.com/rs/416-ZBE-142/images/CrowdFlower_DataScienceReport_2016.pdf

“Data Scientist” Job Postings (2017)

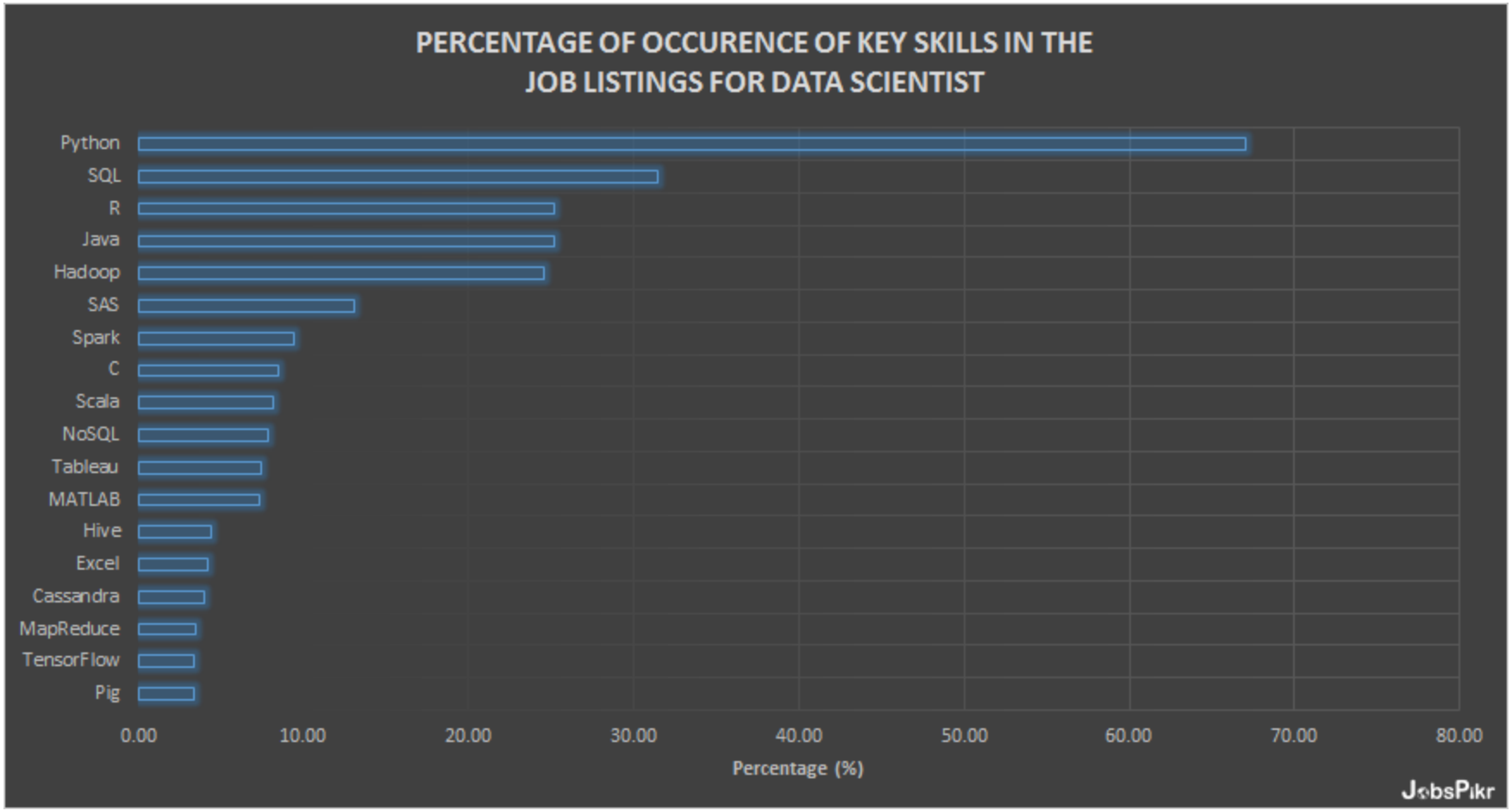
Becoming A Data Scientist: The Skills That Can Make You The Most Money

To pinpoint the most common skills, Glassdoor took 10,000 data scientist job listings that appeared on its job search platform between January and July of this year. The skills required were noted, as were the salaries offered. The data coding skills were extrapolated and analysts searched for those that came up the most within listings. The ten skills that appeared most often as prerequisites for the job, and the percentage of job listings in which they appeared, were:

1. **Python** (72%)
2. **R** (64%)
3. **SQL** (51%)
4. **Hadoop** (39%)
5. **Java** (33%)
6. **SAS** (30%)
7. **Spark** (27%)
8. **Matlab** (20%)
9. **Hive** (17%)
10. **Tableau** (14%)

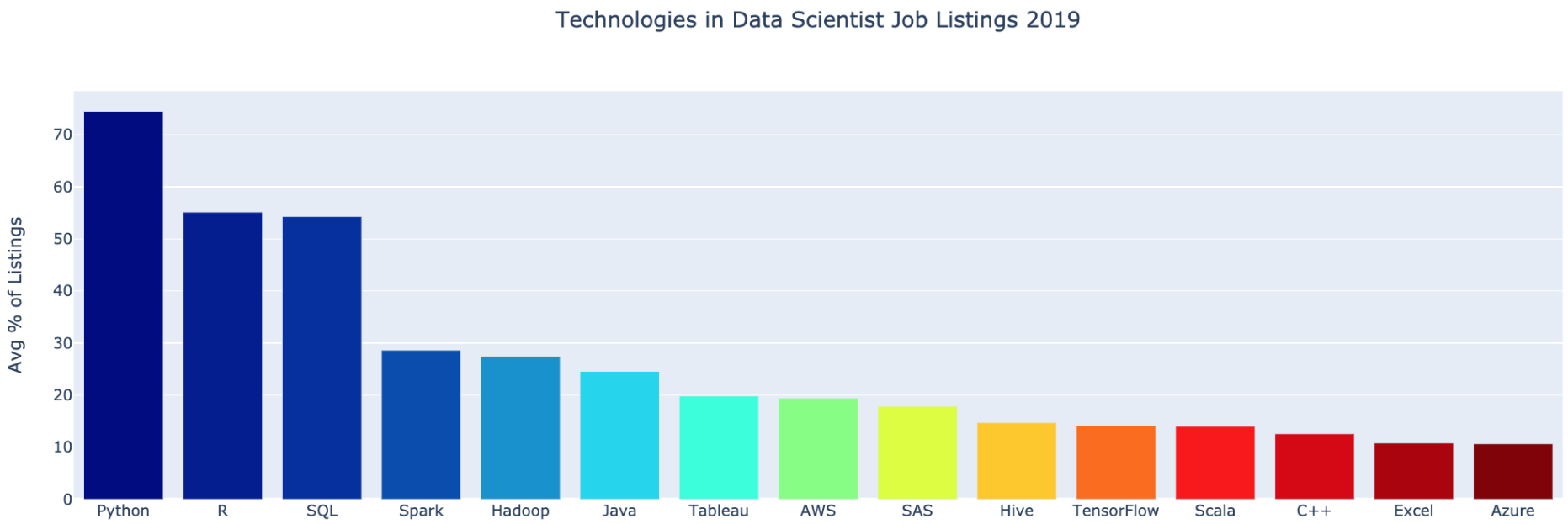
https://www.forbes.com/sites/karstenstrauss/2017/09/21/becoming-a-data-scientist-the-skills-that-can-make-you-the-most-money/

“Data Scientist” Job Postings (2018)



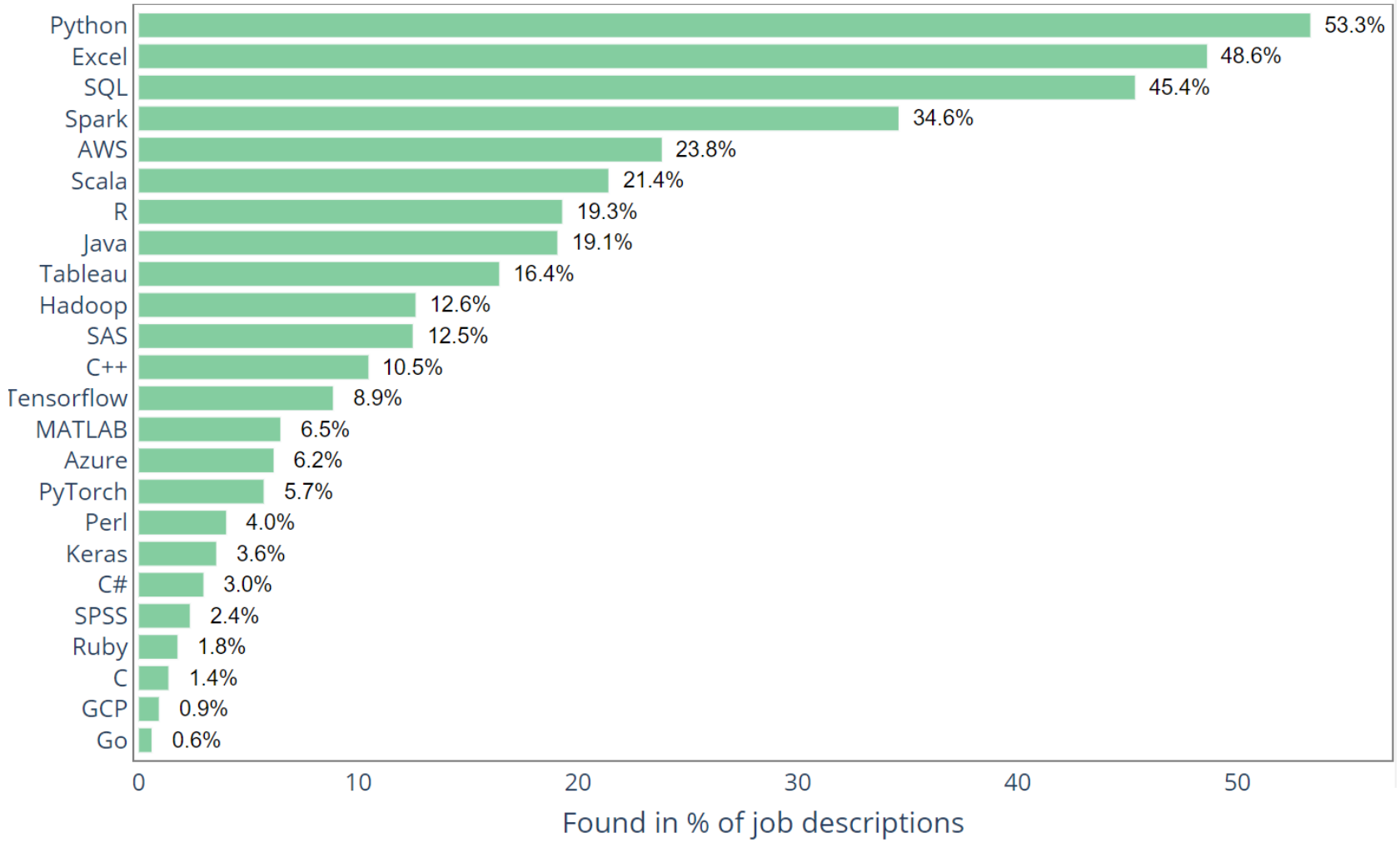
https://www.jobspikr.com/blog/job-data-analysis-reveals-top-skills-required-for-data-science-role/

“Data Scientist” Job Postings (2019)



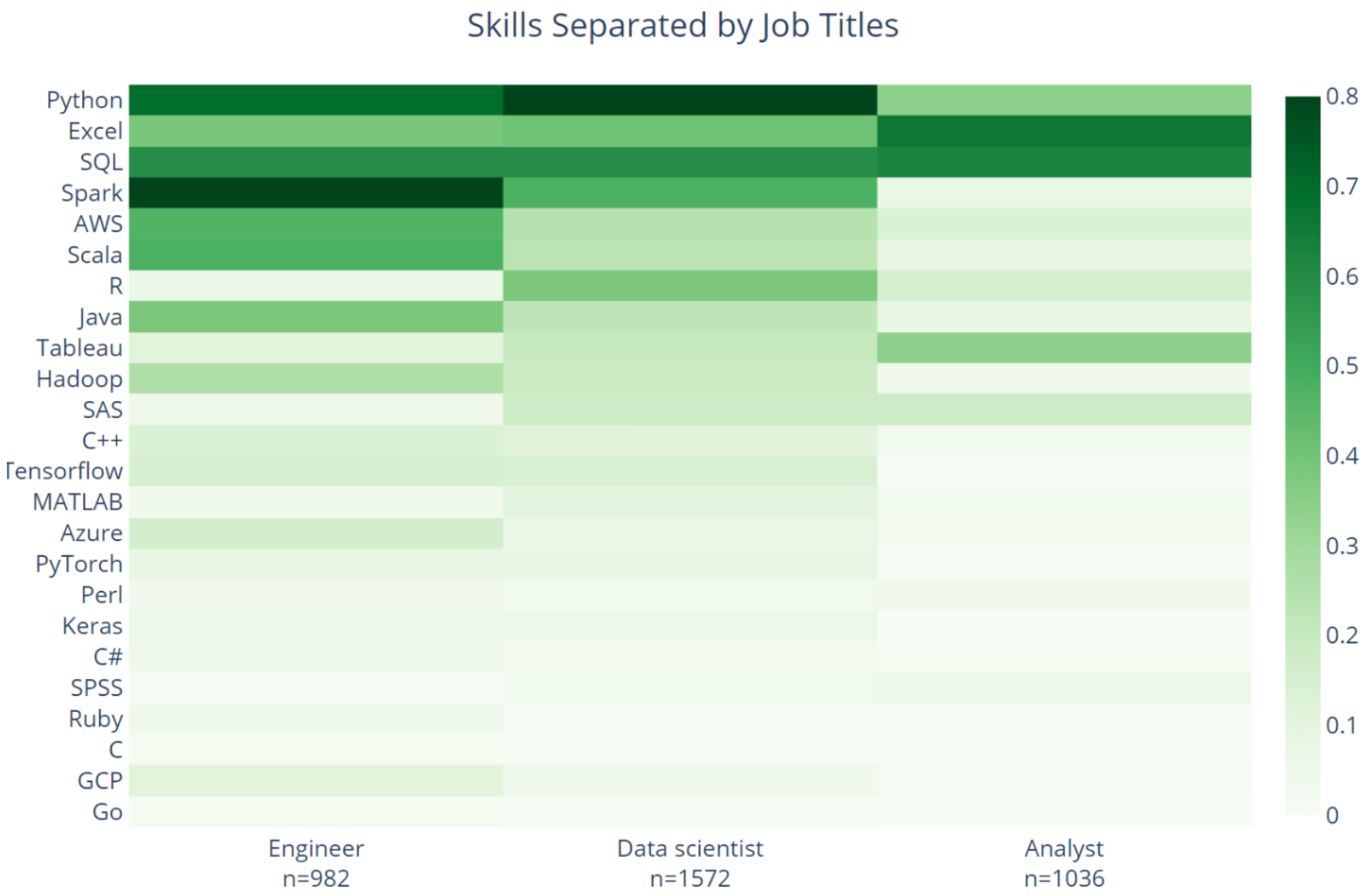
https://towardsdatascience.com/the-most-in-demand-tech-skills-for-data-scientists-d716d10c191d/

“Data Scientist” Job Postings (2020)



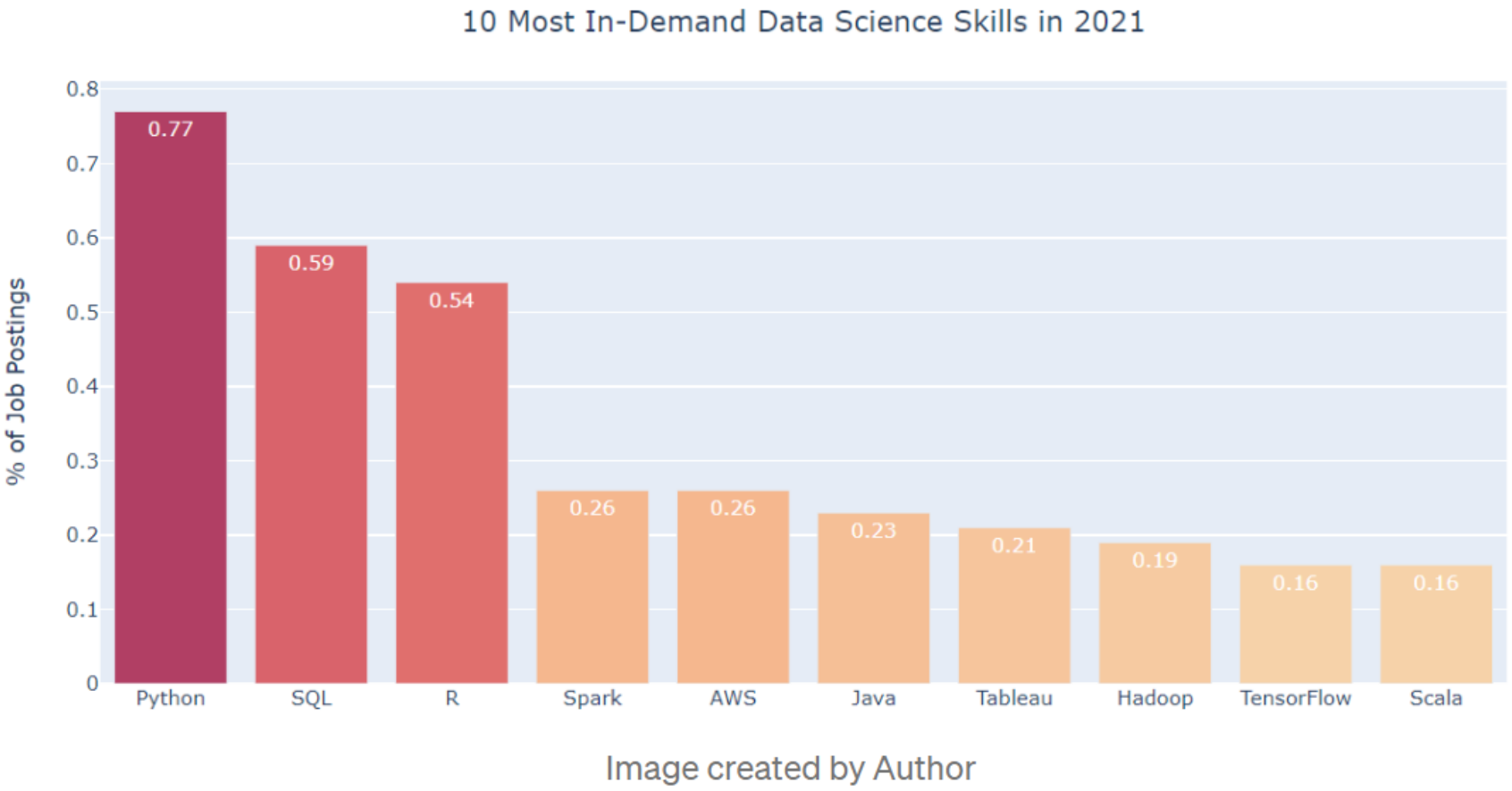
https://medium.com/@chihhsulin/5-500-data-scientist-jobs-report-2020-adefe1d364d3

“Data Scientist” Job Postings (2020)



https://medium.com/@chihhsulin/5-500-data-scientist-jobs-report-2020-adefe1d364d3

“Data Scientist” Job Postings (2021)



Important goal ...

Jobs Companies Degrees

United States / Job / Big Data Consultant

Average Big Data Consultant Salary

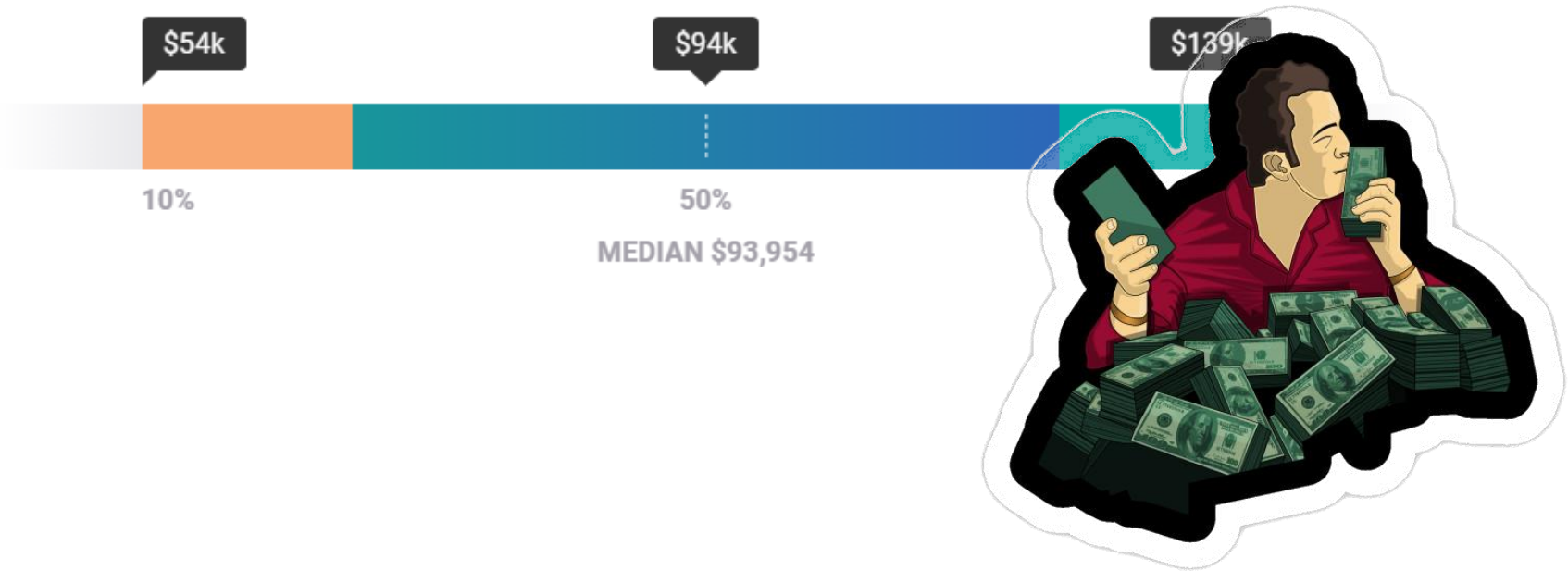
$93,954

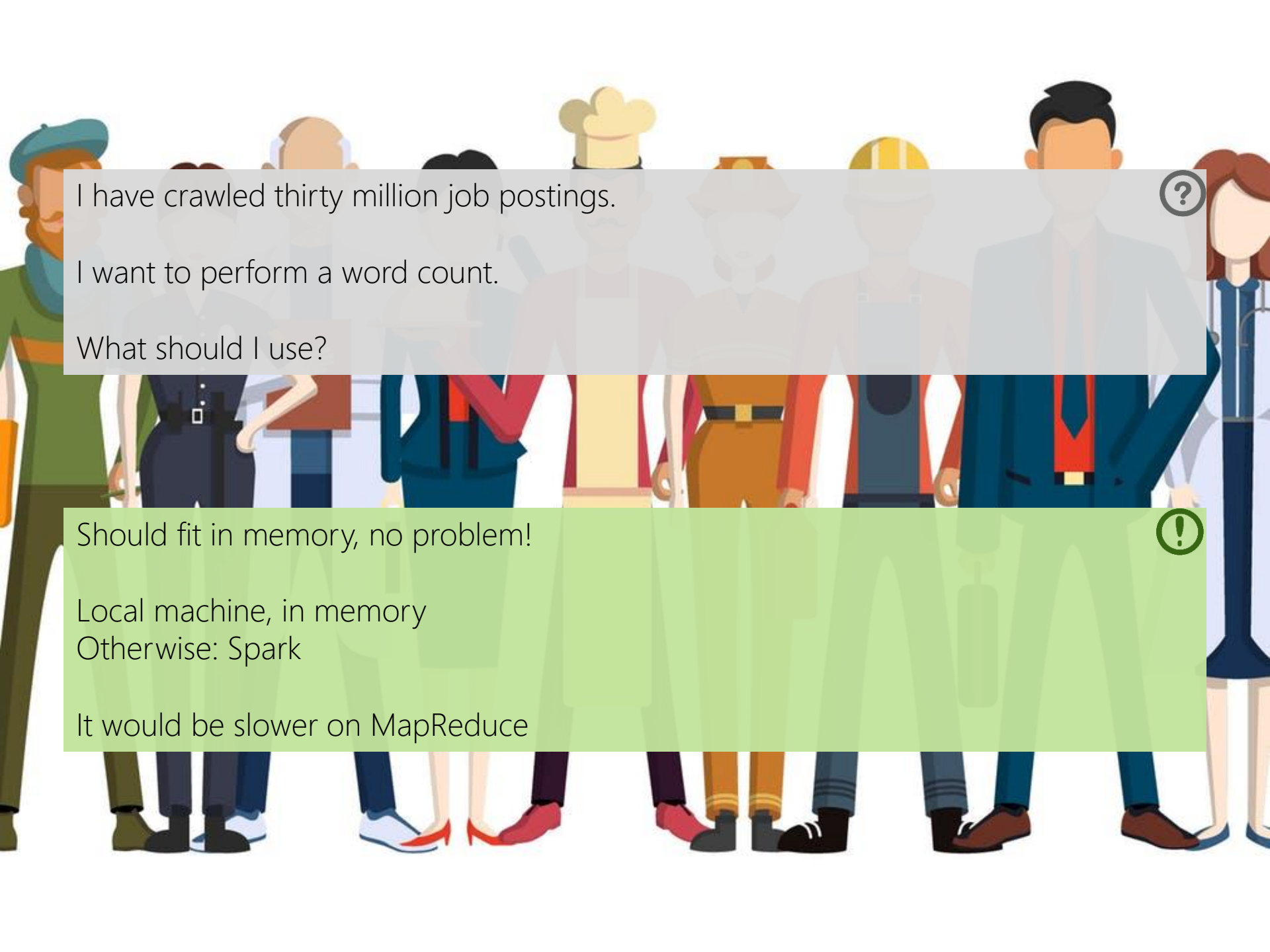
Avg. Salary

$9,826

BONUS

The average salary for a Big Data Consultant is $93,954.





I have crawled thirty million job postings.

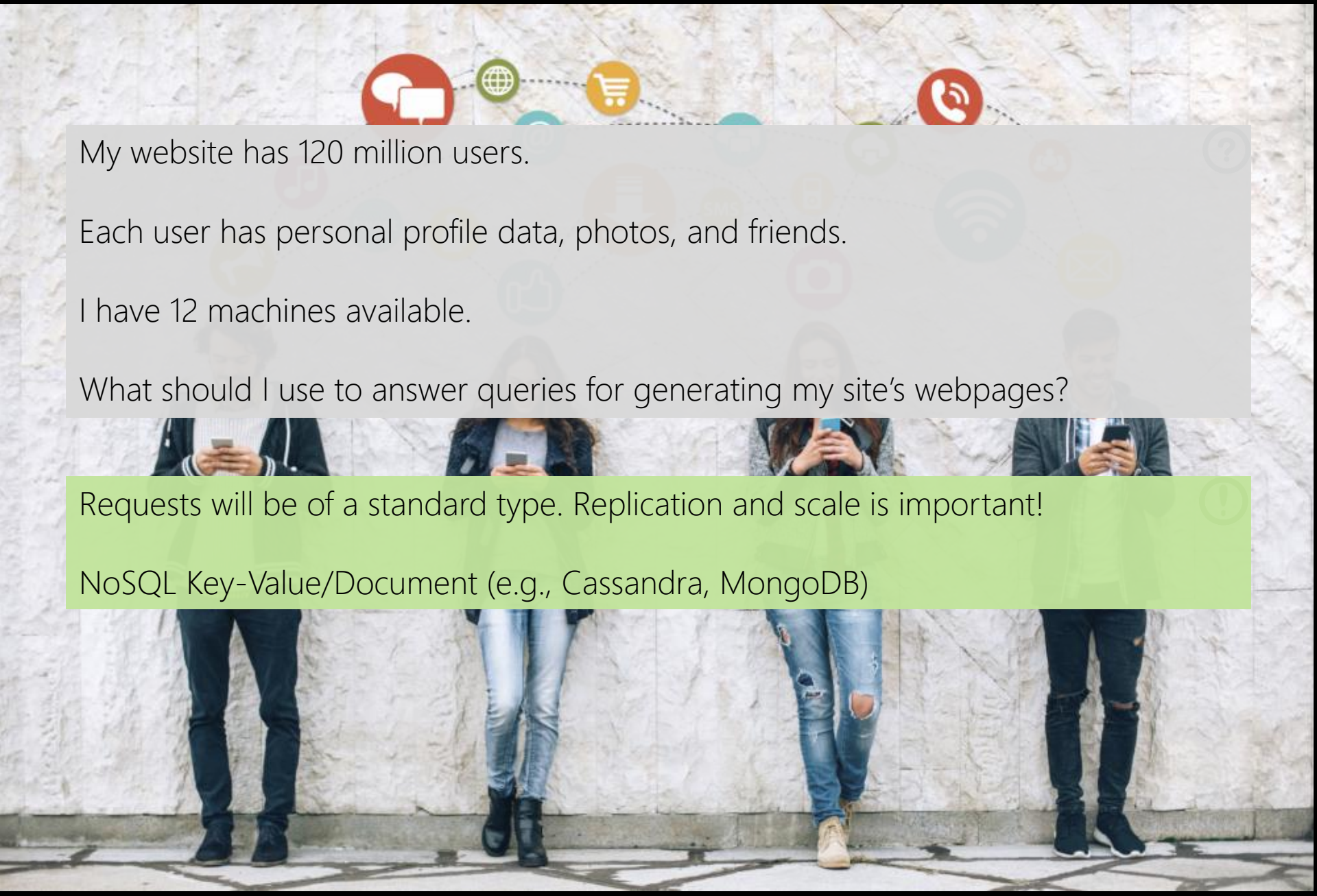
I want to perform a word count.

What should I use?

Should fit in memory, no problem!

Local machine, in memory
Otherwise: Spark

It would be slower on MapReduce



My website has 120 million users.

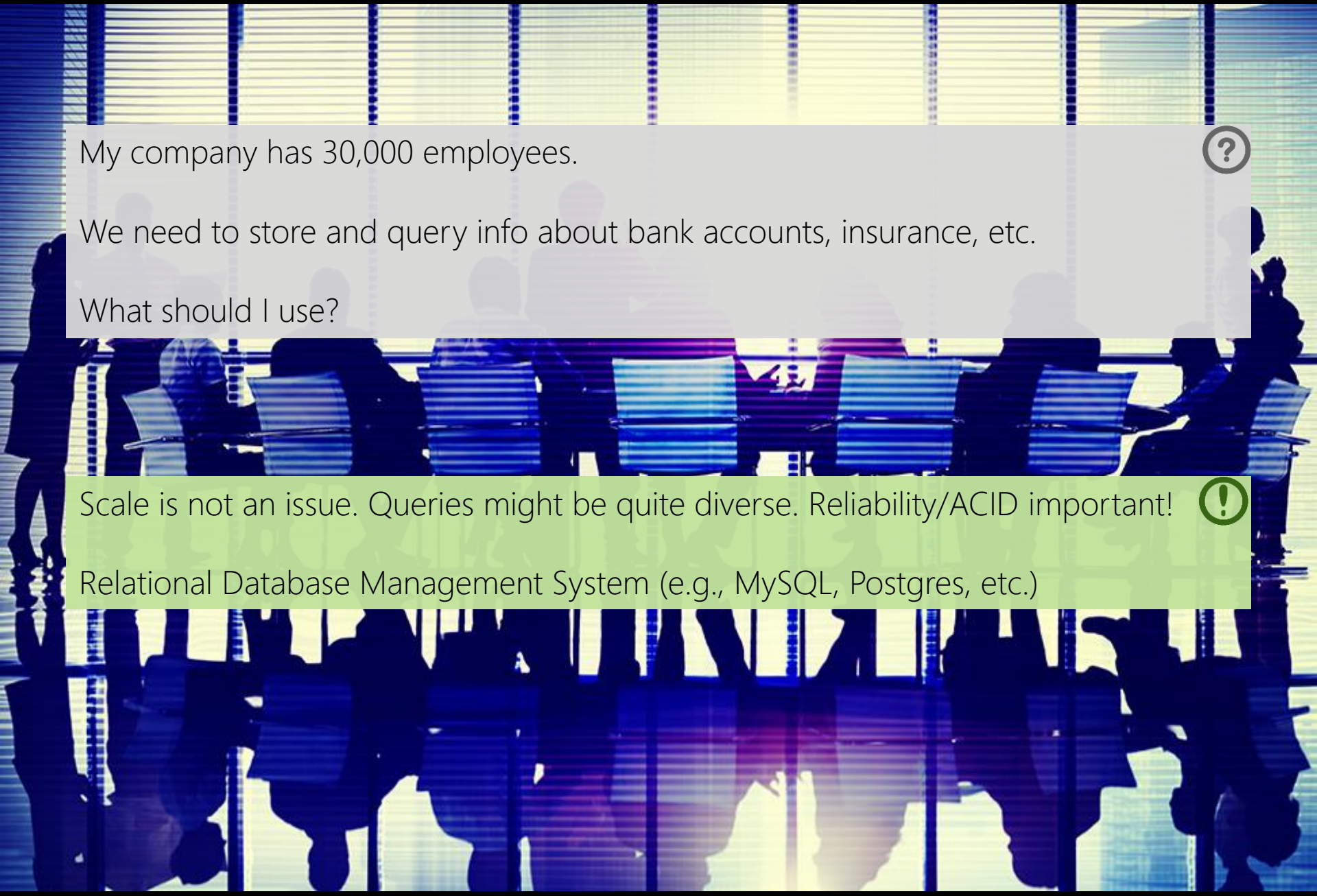
Each user has personal profile data, photos, and friends.

I have 12 machines available.

What should I use to answer queries for generating my site's webpages?

Requests will be of a standard type. Replication and scale is important!

NoSQL Key-Value/Document (e.g., Cassandra, MongoDB)



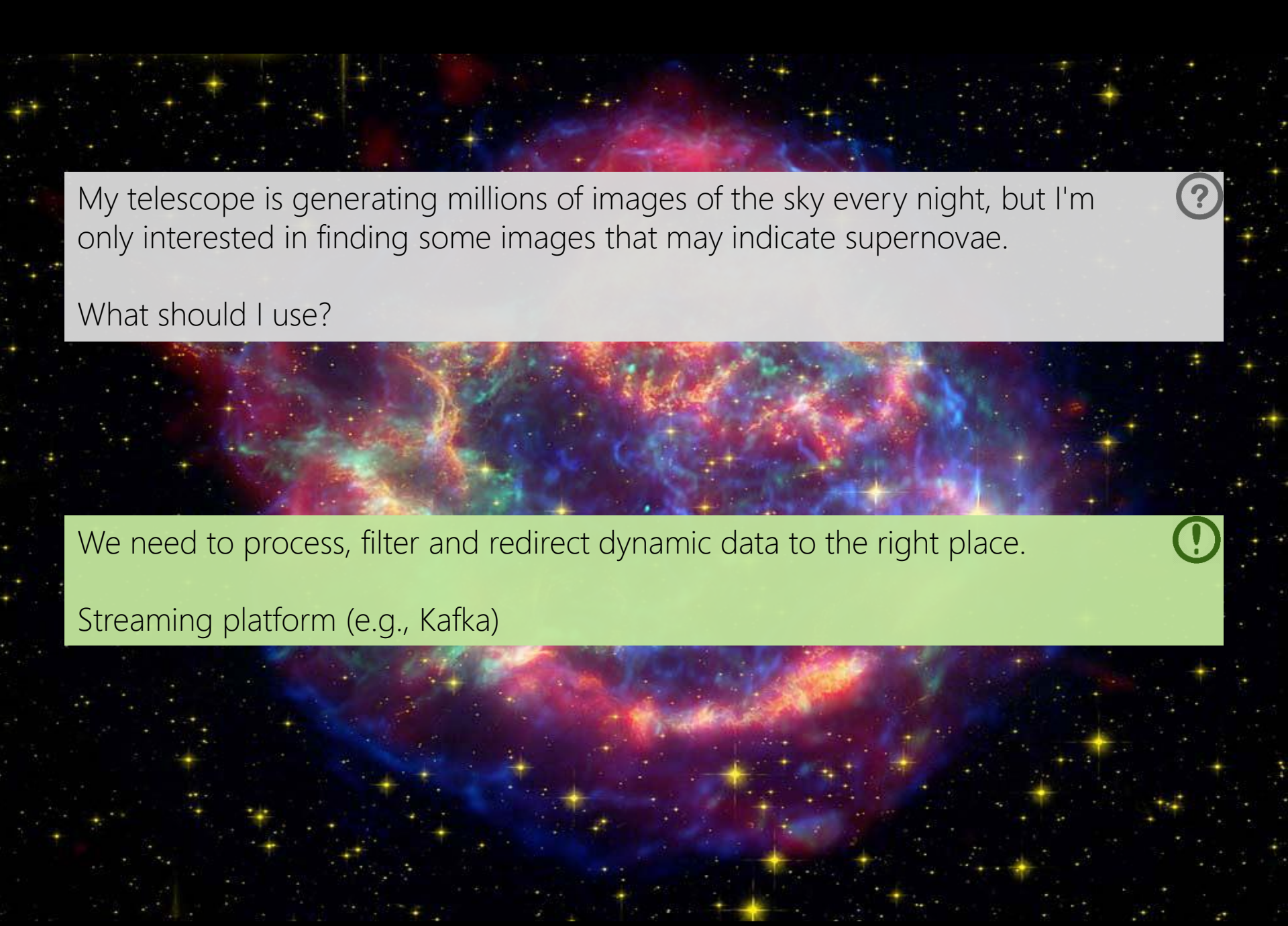
My company has 30,000 employees.

We need to store and query info about bank accounts, insurance, etc.

What should I use?

Scale is not an issue. Queries might be quite diverse. Reliability/ACID important!

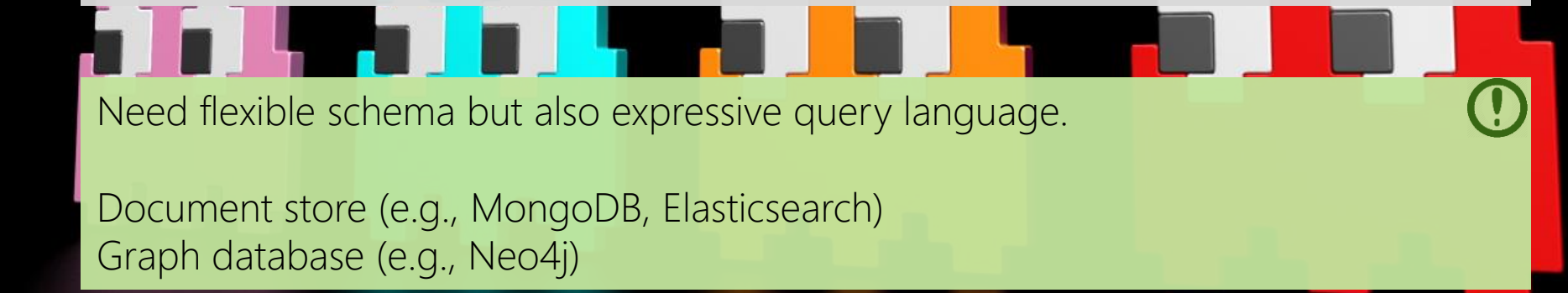
Relational Database Management System (e.g., MySQL, Postgres, etc.)


My telescope is generating millions of images of the sky every night, but I'm only interested in finding some images that may indicate supernovae.
What should I use?
We need to process, filter and redirect dynamic data to the right place.
Streaming platform (e.g., Kafka)

I am scraping data about video games and their characters from various wikis.

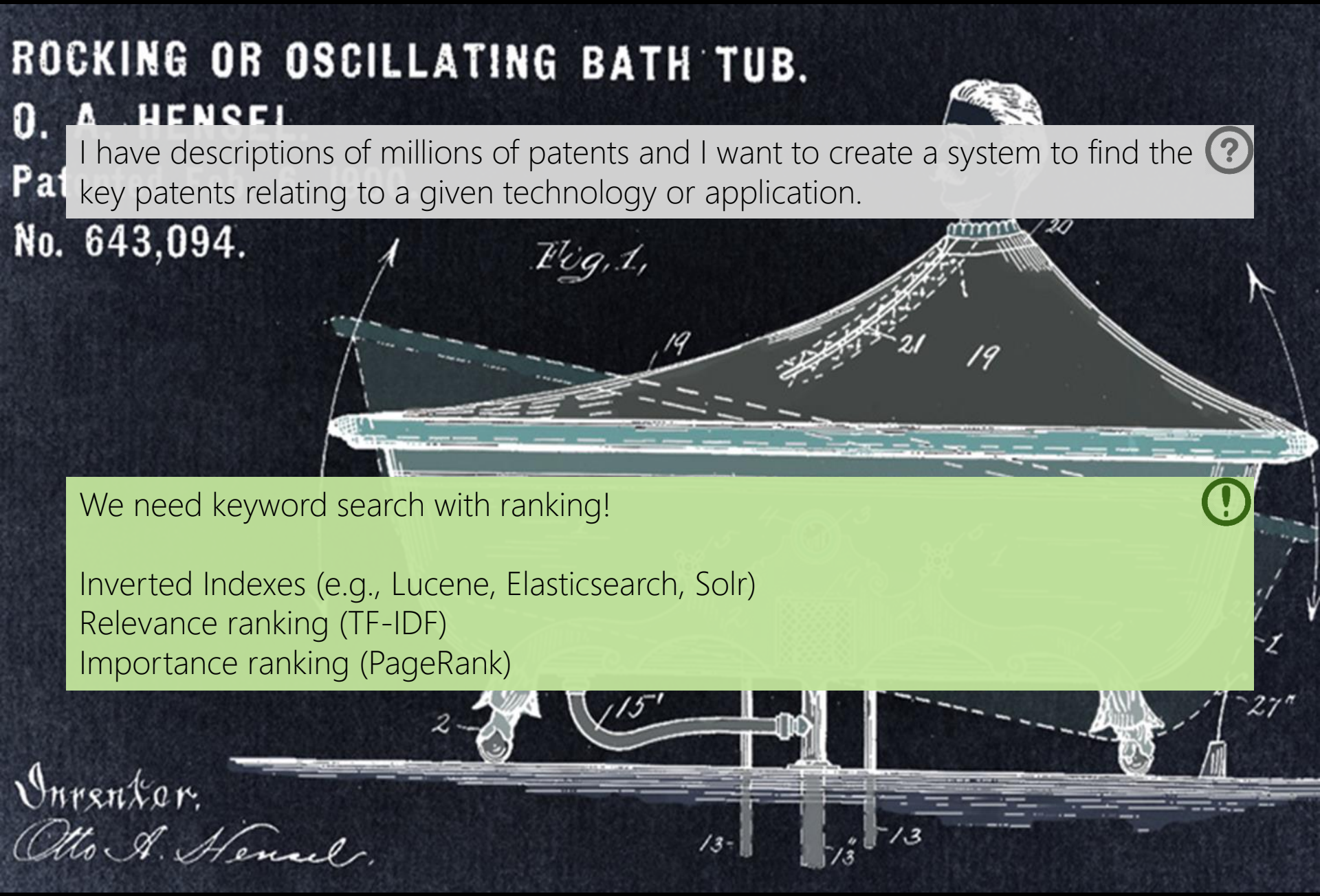
In total I have scraped information from about one million pages and now I want to be able to search over what I have, for example, to find all non-human characters in a particular video game, or platforming games featuring plumbers.

What should I use?



Need flexible schema but also expressive query language.

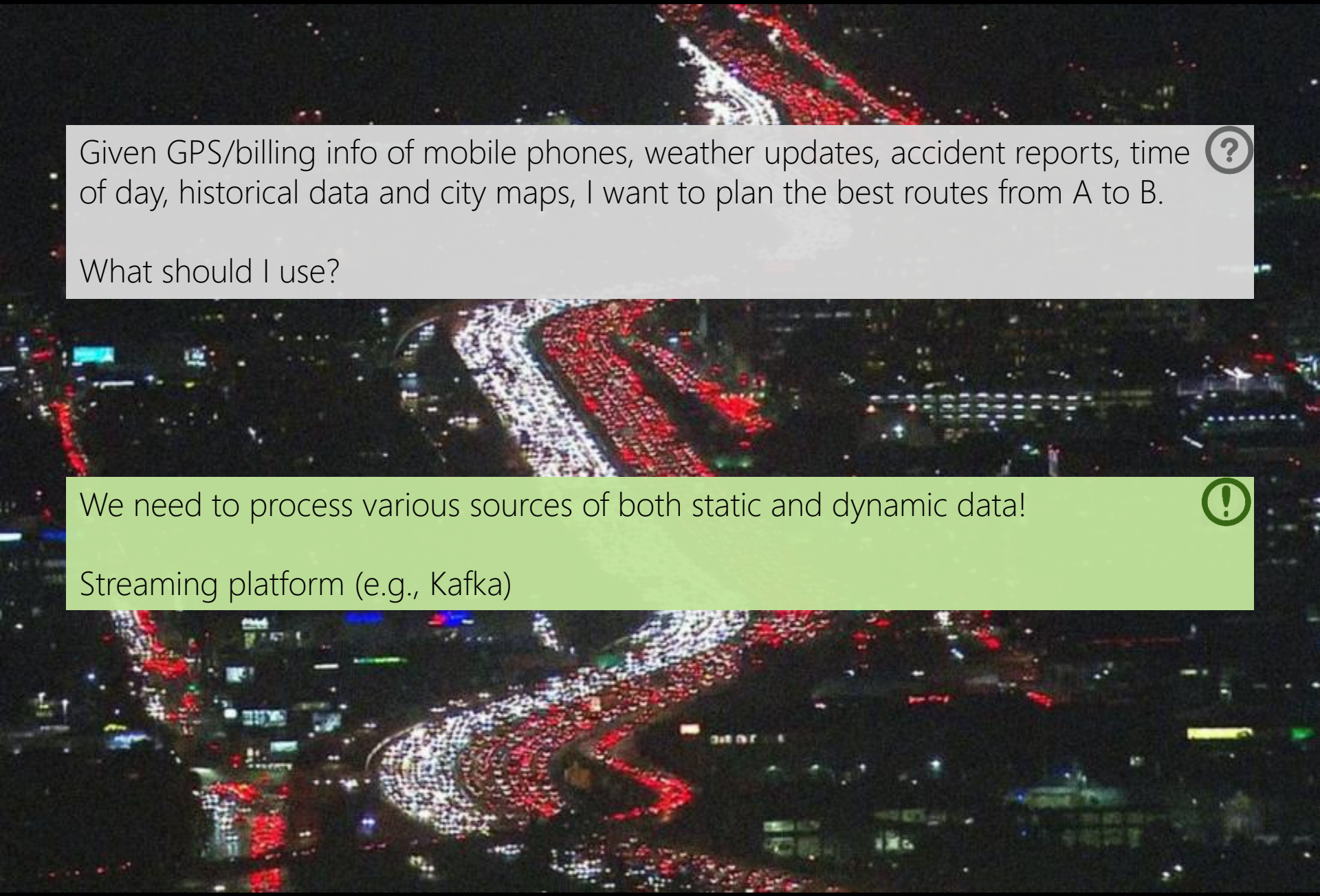
Document store (e.g., MongoDB, Elasticsearch)
Graph database (e.g., Neo4j)



I have descriptions of millions of patents and I want to create a system to find the key patents relating to a given technology or application.

We need keyword search with ranking!

Inverted Indexes (e.g., Lucene, Elasticsearch, Solr)
Relevance ranking (TF-IDF)
Importance ranking (PageRank)

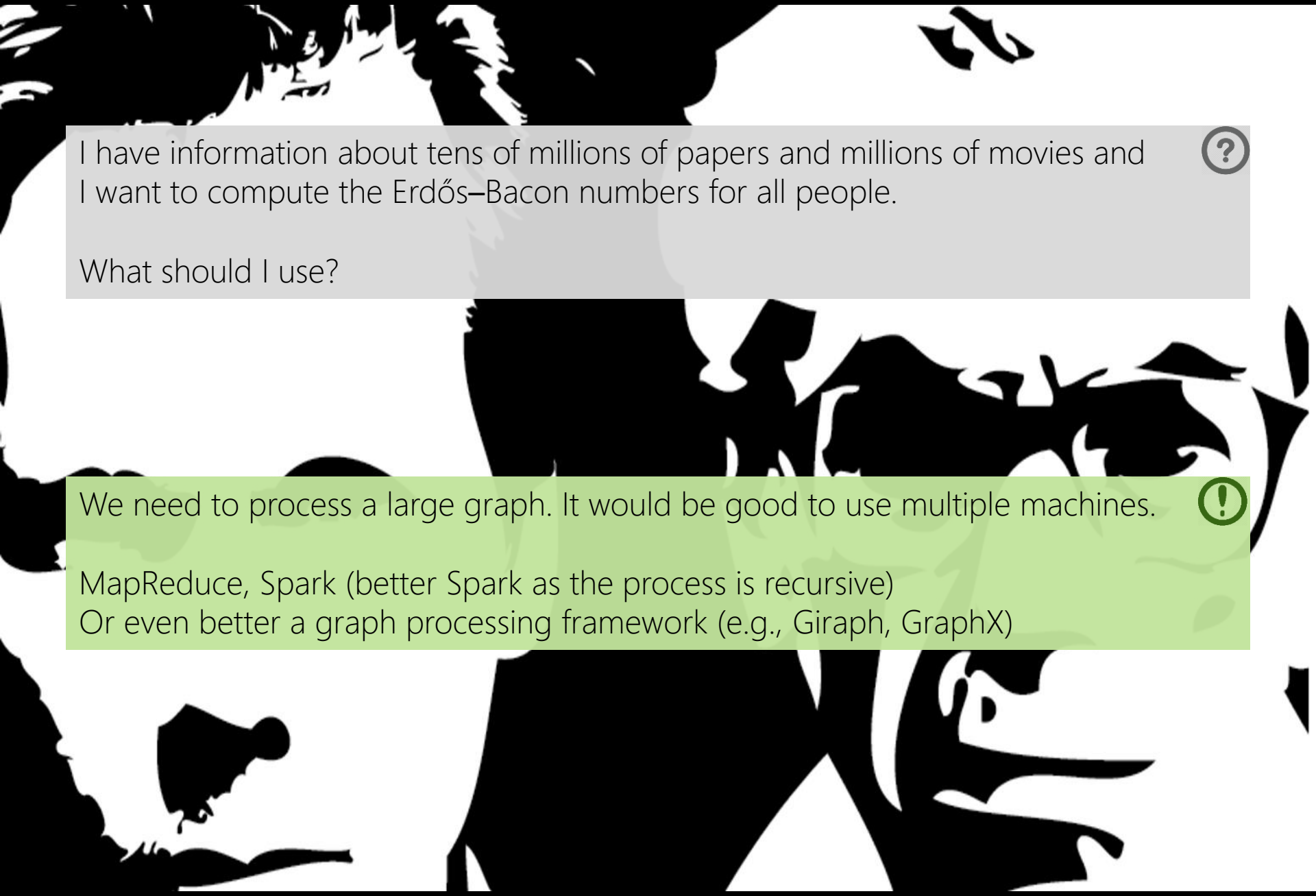


Given GPS/billing info of mobile phones, weather updates, accident reports, time of day, historical data and city maps, I want to plan the best routes from A to B.

What should I use?

We need to process various sources of both static and dynamic data!

Streaming platform (e.g., Kafka)



I have information about tens of millions of papers and millions of movies and I want to compute the Erdős–Bacon numbers for all people.

What should I use?

We need to process a large graph. It would be good to use multiple machines.

MapReduce, Spark (better Spark as the process is recursive)
Or even better a graph processing framework (e.g., Giraph, GraphX)

I am collecting information about research networks in Latin America.

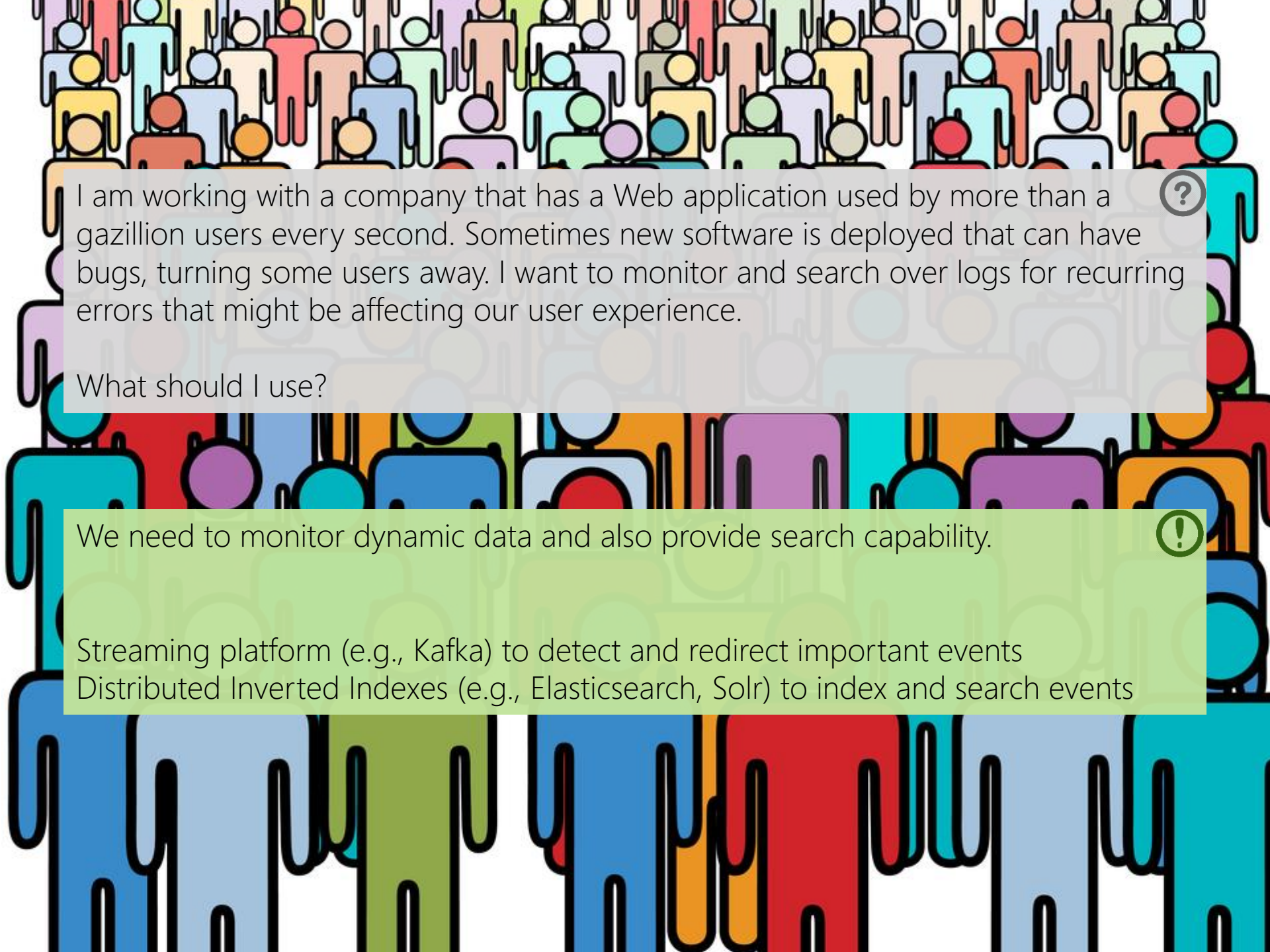
I have information about author affiliations, publications, topics, etc.

Given a particular user, I want to recommend collaborators in the region based on the coauthor network of that user.

What should I use?

Sounds like we might need to search for paths between specific entities.

NoSQL Graph database (e.g., Neo4j)

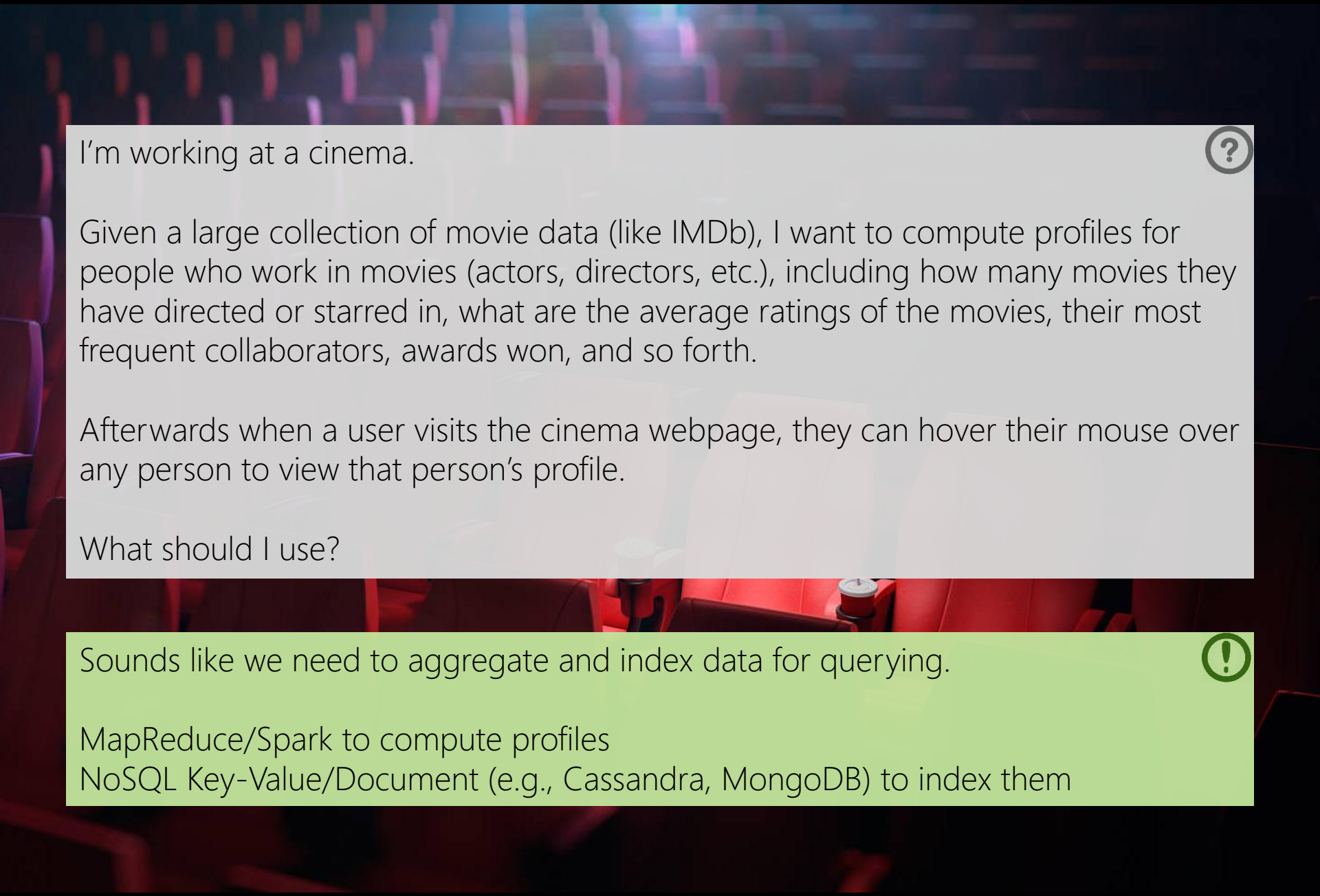


I am working with a company that has a Web application used by more than a gazillion users every second. Sometimes new software is deployed that can have bugs, turning some users away. I want to monitor and search over logs for recurring errors that might be affecting our user experience.

What should I use?

We need to monitor dynamic data and also provide search capability.

Streaming platform (e.g., Kafka) to detect and redirect important events
Distributed Inverted Indexes (e.g., Elasticsearch, Solr) to index and search events



I'm working at a cinema.

Given a large collection of movie data (like IMDb), I want to compute profiles for people who work in movies (actors, directors, etc.), including how many movies they have directed or starred in, what are the average ratings of the movies, their most frequent collaborators, awards won, and so forth.

Afterwards when a user visits the cinema webpage, they can hover their mouse over any person to view that person's profile.

What should I use?

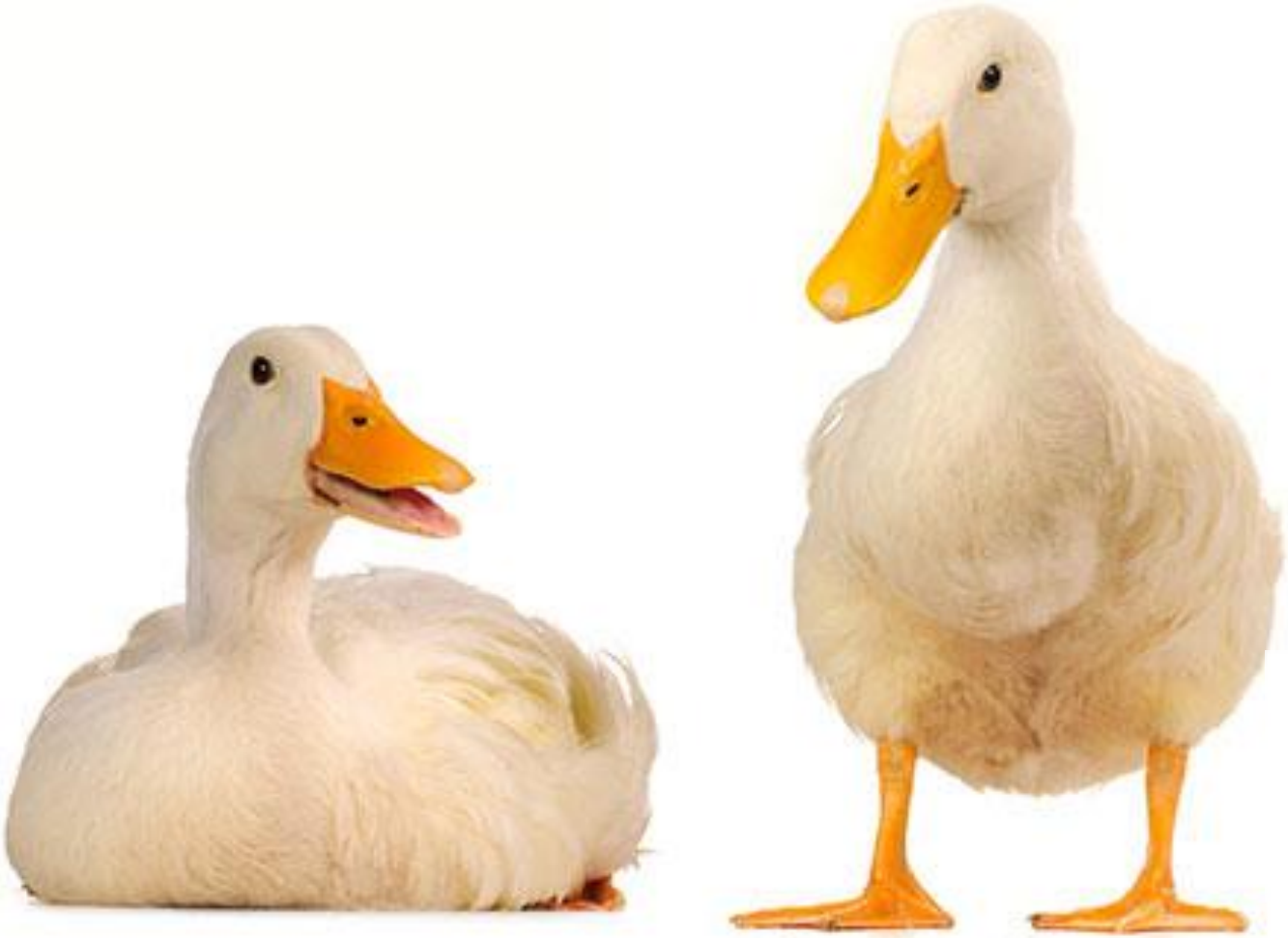
Sounds like we need to aggregate and index data for querying.

MapReduce/Spark to compute profiles
NoSQL Key-Value/Document (e.g., Cassandra, MongoDB) to index them

Final Exam ...

Final Exam

https://pixelastic.github.io/pokemonorbigdata/



Eso.