Business Concept

Overview

Our business goal is to get a better understanding of current technology trends. In this project the main focus will be on the emerging data startups, especially companies related to the modern data stack.

We want to understand which are the hot projects and see how the monitored projects perform compared to others.

The objective was to take a real-life problem and provide a complete, automated and parameterizable solution at the end, where all the stakeholders are represented.

The project will help to provide end-users enabling them to make business decisions, get a real picture of these companies and they can then perform business actions on this basis.

The main objective is to better understand the development of the tech market. To help with this, there are many data sources available and some of which have been selected to help answer these complex questions.

To make this data meaningful for end users, a data visualization tool will be used in the final phase of the project. We want to make the data prepared, accessible and meaningful through dashboards, charts, KPIs and other metrics.

Possible questions the business side may seek answers:

- The active community of open-source and other solutions within a domain
- Rate and popularity of new players compared to old players
- What activity a new release has generated within a given product
- Comparing two projects from the same field by certain attributes

Datasets

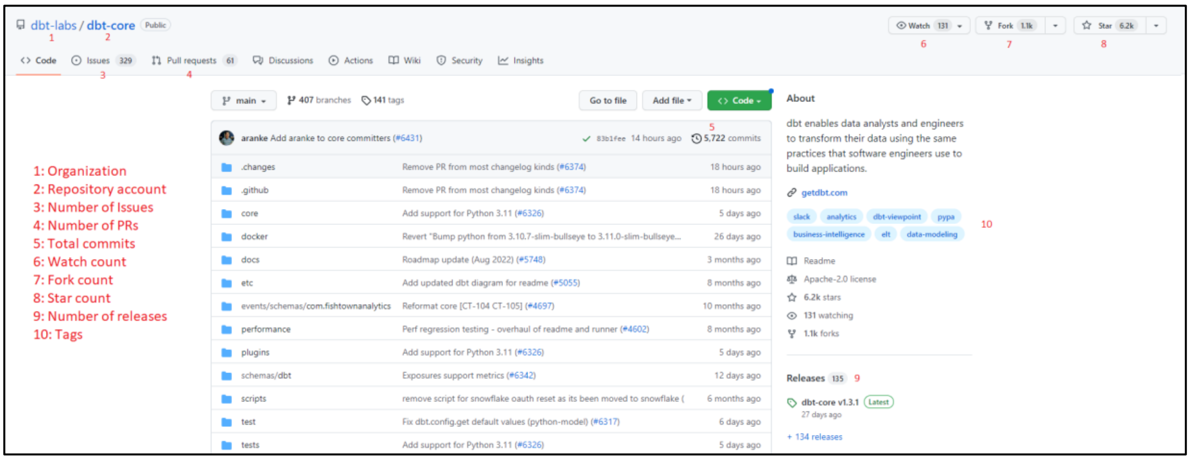
To answer these questions, we will use the following freely available datasets provided by BigQuery:

- GitHub
- Stack Overflow
- Hacker News

GitHub

GitHub is a code hosting platform for version control and collaboration. It lets people work together on projects from anywhere.

The BigQuery dataset comprises the largest released source of GitHub activities. It contains a full snapshot of the content of repositories including PRs, commits, forks and other activities that can be carried out on the site.



1: Organization
2: Repository account
3: Number of Issues
4: Number of PRs
5: Total commits
6: Watch count
7: Fork count
8: Star count
9: Number of releases
10: Tags

aranke Add aranke to core committers (#6431) ✓ 83b1fee 14 hours ago 5,722 commits

.changes	Remove PR from most changelog kinds (#6374)	18 hours ago
.github	Remove PR from most changelog kinds (#6374)	18 hours ago
core	Add support for Python 3.11 (#6326)	5 days ago
docker	Revert "Bump python from 3.10.7-slim-bullseye to 3.11.0-slim-bullseye...	26 days ago
docs	Roadmap update (Aug 2022) (#5748)	3 months ago
etc	Add updated dbt diagram for readme (#5055)	8 months ago
events/schemas/com.fishtownanalytics	Reformat core [CT-104 CT-105] (#4697)	10 months ago
performance	Perf regression testing - overhaul of readme and runner (#4602)	8 months ago
plugins	Add support for Python 3.11 (#6326)	5 days ago
schemas/dbt	Exposures support metrics (#6342)	12 days ago
scripts	remove script for snowflake oauth reset as its been moved to snowflake (	6 months ago
test	Fix dbt.config.get default values (python-model) (#6317)	6 days ago
tests	Add support for Python 3.11 (#6326)	5 days ago

About

dbt enables data analysts and engineers to transform their data using the same practices that software engineers use to build applications.

getdbt.com

slack analytics dbt-viewpoint pypa business-intelligence elt data-modeling

Readme
Apache-2.0 license
6.2k stars
131 watching
1.1k forks

Releases 135

dbt-core v1.3.1 Latest
27 days ago

+ 134 releases

GitHub screenshot with comments

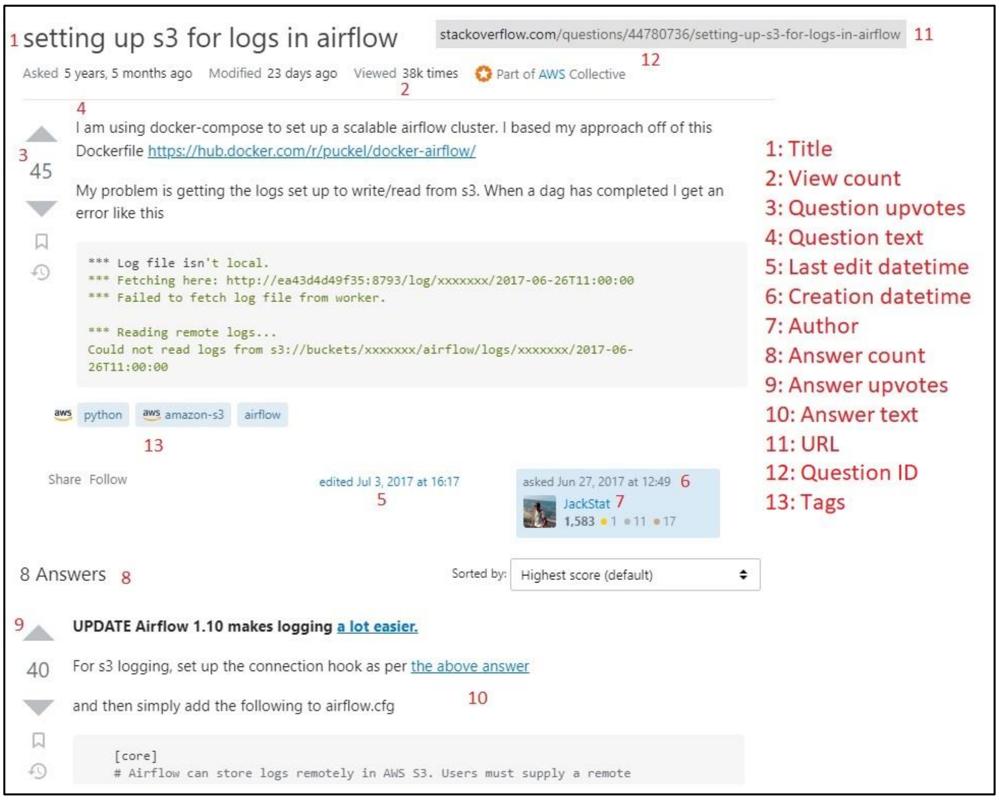
This screenshot shows the home page of a random GitHub repository. The numbers in the screenshot indicate the attributes visible and readable on the web page that are also stored in the BigQuery data source, but in a different format.

For the GitHub data source, the organizing principle for separating the projects will be the organization and possibly an exact repository within an organization.

Stack Overflow

Stack Overflow is a question and answer website for professional and enthusiast programmers. It features questions and answers on a wide range of topics in computer programming. It is the largest online community for programmers to learn, share their knowledge and advance their careers.

The BigQuery dataset includes an archive of all the Stack Overflow content, including posts, votes, tags and badges.



setting up s3 for logs in airflow

stackoverflow.com/questions/44780736/setting-up-s3-for-logs-in-airflow

Asked 5 years, 5 months ago Modified 23 days ago Viewed 38k times Part of AWS Collective

I am using docker-compose to set up a scalable airflow cluster. I based my approach off of this Dockerfile https://hub.docker.com/r/puckel/docker-airflow/

My problem is getting the logs set up to write/read from s3. When a dag has completed I get an error like this

```
*** Log file isn't local.
*** Fetching here: http://ea43d4d49f35:8793/log/xxxxxxx/2017-06-26T11:00:00
*** Failed to fetch log file from worker.

*** Reading remote logs...
Could not read logs from s3://buckets/xxxxxxx/airflow/logs/xxxxxxx/2017-06-
26T11:00:00
```

python amazon-s3 airflow

edited Jul 3, 2017 at 16:17

asked Jun 27, 2017 at 12:49
JackStat
1,583 ● 1 ● 11 ● 17

1: Title
2: View count
3: Question upvotes
4: Question text
5: Last edit datetime
6: Creation datetime
7: Author
8: Answer count
9: Answer upvotes
10: Answer text
11: URL
12: Question ID
13: Tags

8 Answers

Sorted by: Highest score (default)

UPDATE Airflow 1.10 makes logging a lot easier.

For s3 logging, set up the connection hook as per the above answer

and then simply add the following to airflow.cfg

```
[core]
# Airflow can store logs remotely in AWS S3. Users must supply a remote
```

Stack Overflow screenshot with comments

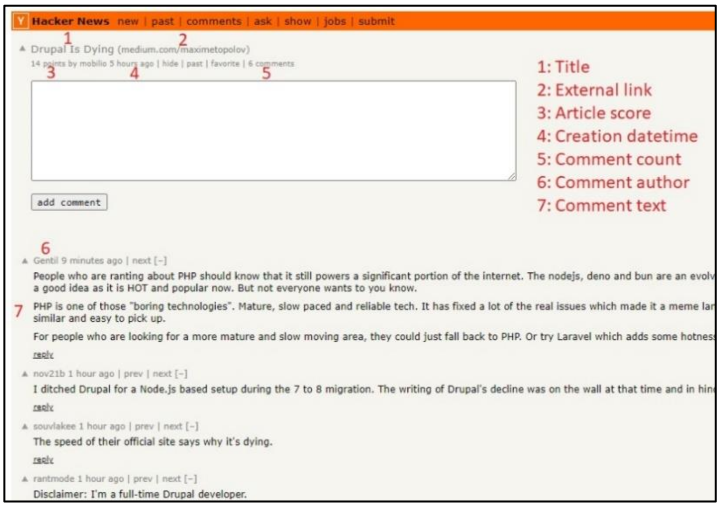
This screenshot above shows a random Stack Overflow question. The numbers in the screenshot indicate the attributes visible and readable on the web page that are also stored in the BigQuery data source, but in a different format.

For the Stack Overflow data source, the organizing principle for separating the projects will be the tags in the questions or possibly a labeling process derived from the question text.

Hacker News

Hacker News is a social news website focusing on computer science and entrepreneurship.

The BigQuery dataset contains all stories and comments from Hacker News from its launch in 2006 to present. Each story contains the author that made the post, when it was written, the number of points the story received and all the related comments.



Drupal Is Dying (medium.com/maximetopolov)

14 points by mobilio 5 hours ago | hide | past | favorite | 6 comments

Gentil 9 minutes ago | next [–]

People who are ranting about PHP should know that it still powers a significant portion of the internet. The nodejs, deno and bun are an evolv
a good idea as it is HOT and popular now. But not everyone wants to you know.

PHP is one of those "boring technologies". Mature, slow paced and reliable tech. It has fixed a lot of the real issues which made it a meme lar
similar and easy to pick up.

For people who are looking for a more mature and slow moving area, they could just fall back to PHP. Or try Laravel which adds some hotness

reply

nov21b 1 hour ago | prev | next [–]

I ditched Drupal for a Node.js based setup during the 7 to 8 migration. The writing of Drupal's decline was on the wall at that time and in hin

reply

souvlakee 1 hour ago | prev | next [–]

The speed of their official site says why it's dying.

reply

rantmode 1 hour ago | prev | next [–]

Disclaimer: I'm a full-time Drupal developer.

Hacker News screenshot with comments

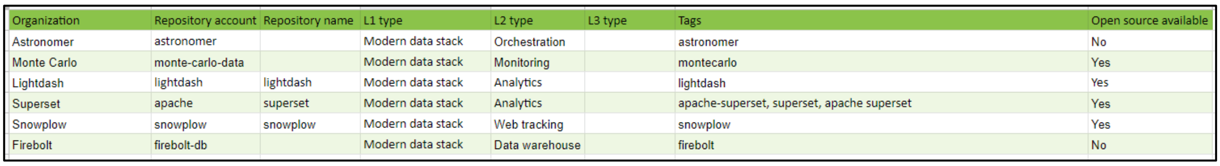
Configuration

From a business point of view, only certain projects are of interest and it is necessary to separate these from the complete data sources.

This requires an input document that contains a list of the companies currently being used, with additional dimensions that are not part of the BigQuery data sources but are of high relevance to the project. For example, what category a company falls into, or whether it has an open-source solution or not.

In this project, it will be stored in a google sheet, which will be ingested and added to our data warehouse.

To simulate real business operations, this google sheet may change from time to time as companies are added or removed along the way. We will also manage these changes as part of the project.



Organization	Repository account	Repository name	L1 type	L2 type	L3 type	Tags	Open source available
Astronomer	astronomer		Modern data stack	Orchestration		astronomer	No
Monte Carlo	monte-carlo-data		Modern data stack	Monitoring		montecarlo	Yes
Lightdash	lightdash	lightdash	Modern data stack	Analytics		lightdash	Yes
Superset	apache	superset	Modern data stack	Analytics		apache-superset, superset, apache superset	Yes
Snowplow	snowplow	snowplow	Modern data stack	Web tracking		snowplow	Yes
Firebolt	firebolt-db		Modern data stack	Data warehouse		firebolt	No

Company details example