Vitor Baptista

✉ vitor@vitorbaptista.com
https://vitorbaptista.com

Experience

2019–   **Data Engineer/Data Scientist**, *Royalty Exchange*, Denver, US.

I'm responsible for the entire data stack, from maintaining the servers (using Docker-Compose, Ansible, GitLab CI/CD), building the pipelines (using Luigi), the data lake (AWS S3), the data warehouse (Snowflake), modeling and transforming the data (using dbt, Data Vault 2.0, and dimensional modeling), building dashboards and reports (using Metabase and Jupyter Notebook), and building machine learning models for entity resolution and time series predictions. Some of my projects were:

- Created pipeline to extracted repertoire data (11M+ artists, 125M+ tracks, compositions, recordings, etc.) from multiple sources, updated monthly
- Built a process to handle entities with multiple identifiers from different sources
- Built a machine learning model for entity resolution of artists
- Analyzed data to understand what affects a song's earnings
- Built a system to continuously monitor the popularity of songs
- Designed our ELT process with a Data Lake and a Data Warehouse

Some of the tools and techniques I used: dbt, Luigi, Snowflake, S3, Ansible, scikit-learn, Jupyter Notebooks, Metabase, Data Vault 2.0, Dimensional Modeling.

2017–2018   **Engineering Lead**, *Open Knowledge International*, Cambridge, UK.

I was responsible for overseeing the software engineering practices across all projects, providing guidance and mentoring to ensure the software we develop is of high quality, on time, and on budget.

2015–2017   **OpenTrials Tech Lead**, *Open Knowledge International*, Cambridge, UK.

I was responsible for the technical aspects of the OpenTrials (https://opentrials.net) project, which aggregates clinical trial data from government registries across the globe. The team was formed by three developers, one project manager, and one community manager.

2012–2014   **Developer**, *Open Knowledge Foundation*, Cambridge, UK.

I was part of the CKAN (http://ckan.org) team. CKAN is the world's most used open source data portal portal, powering sites like http://data.gov.uk and http://data.gov. It's written in Python using Pylons and Postgres. The core team was composed of about 10 people geographically distributed.

My latest contribution was revamping its data visualization system, allowing a single dataset to have multiple visualizations, and making it easier to build custom visualizations. This feature was initially developed for London's Natural History Museum Data Portal (http://data.nhm.ac.uk/), and was later released as part of CKAN 2.4.

I've also deployed a few data portals like http://datos.gob.mx and http://data.org.pk, making sure their servers are secured and able to handle their expected load. customizing them based on the clients' requirements.

2011–2012 **Developer**, *ThoughtWorks*, Porto Alegre, Brazil.

I've worked as a Rails developer for a few projects during my time at ThoughtWorks. The projects were managed following agile development practices, including pair programming.

My main contribution was being part of the team of 6 people that flew from Brazil to the USA to work on-site for 3 months with one of the (at the time) largest clients of ThoughtWorks. The objective was to meet the client and their internal developers, work with them, understand the code structure and their challenges, to later bring the project to Brazil.

It was one of the largest and oldest Rails codebases worldwide, with about 300k lines of code, in production since 2007. The team had 50+ people distributed between two locations in the USA, Canada, India, and Brazil.

2010 **Intern**, *Linux Foundation*.

I was selected as a student in Google's Summer of Code 2010 for the Linux Foundation. I've worked for 3 months with the leader of the OpenPrinting project, Till Kamppeter, developing in Python a compressor for PostScript Description Files (printer drivers) named pyppd (`https://github.com/OpenPrinting/pyppd`). I was able to get 90% of compression ratio. It's included by default in all the main GNU/Linux distributions, like Ubuntu, Debian and Fedora.

Education

2013–2015 **Master's in Computer Science**, *Federal University of Paraíba (UFPB)*, Paraíba, Brazil.

I've built a statistical model using R and Python that detects when a legislator is changing her position in relation to the government, entering or leaving its coalition, based on her voting patterns during rollcalls. The final model was based on the C5.0 method. It achieved 90% accuracy with an area under the ROC curve of 0.88.

2006–2010 **Bachelor in Computer Science**, *Federal University of Paraíba (UFPB)*, Paraíba, Brazil.

I've built a software that returns if two or more software licenses are compatible or not. They have to be described in the Creative Commons Rights Expression Language (ccREL). This could help companies and developers to better understand if they can add in their projects a third-party library that's licensed under some other license.

Projects

Queremos Saber
To make a Freedom of Information Request in Brazil you need to provide your identity. This has led to some people receiving threats, and FOIA requests being treated differently depending on who was asking (e.g., delaying answers to known journalists). To solve this, I built and maintain Queremos Samber (`https://queremossaber.org.br`), a platform where any person can anonymously submit a FOIA request for any public body, which will then be sent manually by us under Open Knowledge Brazil's name, protecting the requestor identity. This circumvents the law's requirements while preserving the citizen's identity.

ShellShare
I built and maintain a platform that allows sharing your terminal session live with a one-liner command. It works on Linux, Mac, and Windows. It uses Python for the client and NodeJS for the server. It has been used in classroom trainings. (`https://shellshare.net`)

Who won 2012's elections in Brazil?
I've worked with Estadão's data visualization team during the week before municipal/state elections in Brazil. We designed and built an interactive visualization that shows the growth and decrease of the political parties in the states. How many mayors they got elected, how many people they'll govern, etc.. It's a single-page app built with D3.js. We updated the data live, while the results were being published. There were almost 150.000 visits in two days. (`http://estadaodados.com/eleicoes2012/`)