Dandelion

Software Engineer (Big Data and ML Infrastructure)

New York or Remote / Full-Time / Mid-Career / Immediate Start

Resumes, questions, and requests for assistance or an accommodation due to a disability may be directed to recruiting@dandelionhealth.ai.

Salary Range: $150K - 175K

Our Team

Dandelion Health was founded in 2020 by experts in health tech, hospital systems, academia, and clinical AI. We are building the world's largest AI training and validation platform. Today, we pride ourselves on our ability to make data access as easy as possible for AI developers, while raising the bar for patient safety and data quality. Tomorrow, we will be the place where any AI developer can go to build a responsible clinical AI product. Our culture is all about learning from data and improving, so we can help our clients improve health through AI. Meet the rest of our team here.

Your Role

Dandelion works with healthcare data in the petabyte range, so having world-class data architecture and stewardship is critical to our shared success. You will work closely with Technical Product Management, Data Science, and Health System Partners to develop scalable solutions to process and store terabytes of structured and unstructured data from legacy systems to a cutting-edge machine learning platform. You should have deep expertise building data platforms on AWS, Azure and/or GCP. You should have excellent business and interpersonal skills to work with internal and external stakeholders to understand data requirements to implement efficient and scalable ETL solutions. You should be comfortable working in ambiguity, and embody perseverance and practical problem solving.

Responsibilities

- Design, implement, and manage data pipelines between health systems and Dandelion's data platform, with a focus on transparency and auditability;
- Partner with clinical informaticists and data analysts to surface data, process, regulatory, and technology issues through identification, measurement, and monitoring of our operations;
- Become a trusted partner to health system stakeholders, allowing the movement of up to 25 PB of clinical data from legacy healthcare systems into a secure Cloud environment;
- Develop and improve existing ETL infrastructure and tooling with an emphasis on data quality, efficiency, and security.
- Summarize the complexity of these data pipelines and operations into clear explanations and documentation for internal and external audiences.

Qualifications

Required technical skills

- 3+ years of Python development experience across the full software development lifecycle (design, implementation, testing, deployment, maintenance)
- Experience working with OLAPs (i.e. AWS Redshift, Google Bigquery, Snowflake) and comfort with SQL
- Ability to create and maintain cloud infrastructure components via Terraform or other infrastructure as code in AWS, GCP, or Azure
- Experience using virtualization and containerization (e.g. Docker, VMware)
- Experience interacting with business users to determine optimal infrastructure and deploy software solutions
- Proficiency with one or more command languages (e.g. Bash)

Required non-technical skills

- 2+ years of work experience with non-technical stakeholders
- Experience designing and improving workflows and standing up accompanying operating and technical procedures
- Strong analytical decision-making and organizational skills
- Perseverance and practical problem solving
- Humility and strong team collaboration

Preferred

- Experience with healthcare data (familiar with HIPAA PHI elements)
- Experience with designing and implementing ETL pipelines at scale
- Experience with data quality, QA, and validation
- Familiarity with NLP, OCR, and/or ML tools, especially in Python or cloud services
- Startup experience
- Proficiency in data harmonization, database architecture, and/or cloud computing