Arpit Solanki

Data Engineer

☎ +91-8238059340 | ✉ solankiarpit1997@gmail.com | ⌂ arpitsolanki.in | arpit1997 | arpit-solanki | in arpit-solanki

Education

Indian Institute of Information Technology *Vadodara, IN*

Bachelor of Technology in Computer Science, CGPA: 7.8/10 *May 2018*

Experience

Atlan (formerly SocialCops) *Delhi, IN*

Data Engineer *Jun. 2018 - Present*

- Built data quality as a service using Spark to calculate metrics such as mean, median, PII columns, outlier percentage, nulls whilst providing an interface to implement custom metrics by the user and supporting data sources like Snowflake, Data Lakes on S3, Oracle. User can also setup tests on data like mean(columnA)>20 and service reports passed and failed checks to the user. This brought trust into the data for the user.
- Built an AWS deployment of open source presto with high available presto coordinator nodes, graceful shutdown of workers nodes to support autoscaling, Cloudwatch agent to push metrics like cpu, heap usage to support monitoring and autoscaling to power ETL and itanteractive query workloads.
- Built a micro service in Java to power objects and data level authorisation like masking, column level access control for SQL queries in AWS Athena. I built this using ANTLR grammar of presto and query rewrite methods.
- Added features like data deduplication, data extraction from SQL file dumps, data ingestion from S3, running SQL queries in the workflow to Workflows component of the product
- Improved cataloging performance of Catalog component of the product by 90% using Presto and Hive. Improved the data versioning by reducing the storage used by over 80%
- Built data pipelines using Airflow for Ministry of Rural Development India's project DISHA which includes data ingestion from third party APIs, transformation using Python/R and data sink Elasticsearch including data versioning to rollback to previous versions.
- Managed on-premise deployments for DISHA project, packaged the entire stack of Vue.js frontend app, Node.js backend and Airflow pipelines into Debian package.
- Created a health monitoring service for DISHA deployment using Prometheus and Grafana to alert stakeholders of critical scenarios like failing ETL pipelines, high resource usage, high traffic on dashboard etc.

Codementor *Remote*

Mentor *Jan. 2019 - Present*

- Took 100+ sessions with average 5 star rating and mentored developers in technologies like Python, Airflow, Spark and assisted them in problem solving

SocialCops *Delhi, IN*

Data Engineer Intern *Jan. 2018 - May 2018*

- Built a data integration framework using celery with support for state management and incremental data ingestion.
- Worked with data scientists and Bussiness intelligence team to build ETL pipelines using Airflow to power analytics and dashboards
- Worked on building system native packages (Debian and RPM) of data integration services and ETL pipelines written in Python.

Projects

Human Detection and Tracking *Vaodara, IN*

Project Lead *June 2016*

- A command line application built with OpenCV and Python to detect individual humans and track them using their faces
- The application provides an API to users to train their own face datasets, the feature was built to support use cases like surveillance systems. The project has got 600+ stars, available on Github.

Skills

Languages	Python, Scala, Java, Shell scripting
Big Data Tools	Presto, Spark, Hive, Airflow, Kafka, Elasticsearch, Delta Lake Parquet
Tools/Platform	Kubernetes, AWS, Prometheus, Grafana, Docker, Celery

Others

Stackoverflow - Written 230+ answers, overall top 5% in Python with 5.7K reputation

Blogs - Testing Celery Tasks - `https://medium.com/@solankiarpit1997/testing-celery-tasks-a8a3568d0213`