Arun Krishnan

Krishnankripa, Vilakkupara, Anchal, Kollam, Kerala PIN-691312

☎+91-8086376896 ✉ arun19krishnan96@gmail.com in linkedin.com/in/arun-krishnan-08661885

Summery

With a solid foundation in Data governance, I have successfully transitioned into a Data Engineer role, specialising in ETL processes, database modelling, and optimising data pipelines for smoother data flow. I have a track record of enhancing data pipeline efficiency by as much as 30% and cutting processing time by 20%. Additionally, I have extensive in building web crawlers and overcoming blocking mechanisms. I am committed to leveraging data-driven insights to improve data management strategies.

Technical Skills

- **Programming Languages:** Python (Pandas, Numpy), SQL
- **Cloud & Devops**: Azure Data Services, Docker, CI/CD, Kubernetes
- **Machine Learning & Data Science:** Machine Learning, Scikit-Learn
- **Etl & Data Pipelines:** Apache Airflow, Apache Kafka, Azure Data Factory
- **Testing & Data Quality Assurance:** Unit Testing, Great Expectations(Ge)
- **Databases & Storage:** Postgresql, Aws S3, Elasticsearch, Azure Data Lake Gen2
- **Big Data & Processing:** Pyspark, Apache Flink, Nessie, Apache Hive, Azure Databricks

Experience

Turbolab Technologies Pvt. Ltd. (Scrapehero) July 2022 - Present

Associate Data Engineer *Kochi*

- Architecting scalable data infrastructures using PySpark, Dremio, and Iceberg to handle large volumes of data, achieving a scalability improvement of up to 40%.
- Designing and implementing efficient ETL processes, reducing data processing time by 25% and enhancing data quality by 20%
- Developing and optimising data models and schemas using dimensional modelling for effective data storage and retrieval, improving data access speed by 30%.
- Collaborated with cross-functional teams to integrate machine learning models into the data pipeline, increasing predictive analytics accuracy by 25%.
- Automated data quality checks, reducing data processing errors by 20% and ensuring high-quality data for analysis.
- Implementing data governance policies and procedures to ensure data security, compliance, and privacy, achieving a 95% data accuracy rate.
- Implement Great Expectations(GE) to improve data quality by 67%, utilising data profiling, validation, and automated monitoring to ensure adherence to essential quality metrics like completeness, uniqueness, and format accuracy.
- Managed comprehensive data quality workflows with Apache Airflow, incorporating Great Expectations for data validation and monitoring. This approach enhanced data quality by 67% and decreased pipeline failures by 30% through proactive detection and resolution of issues.
- Developed robust web crawlers in Python for e-commerce websites. Achieved a 40% increase in data extraction efficiency through optimised web crawlers.
- Organised and stored scraped information into databases for easy accessibility. Improved data storage and accessibility by 30% through efficient database management techniques.
- Bypassing blocking measures, such as CAPTCHA challenges, IP blocking, and user-agent detection.
- We are using Kubernetes to deploy the pipeline, which will automate the deployment process and make it 95% easier.

Core Competencies

- Analytical
- Innovative
- Strategic
- Decisiveness
- Mentoring
- Debugging
- Research
- Documentation

Projects

- E-commerce Data Streaming ETL Pipeline (Professional)
  - Designed a scalable architecture of a pipeline that consume data from Kafka(source) using PySaprk as RDD, it improve 45% faster processing capability compared to normal data frame. Then Transform the data for compact to the iceberg and Postgres table, then process the data under quality validation so it can improve data quality by 67%. The it write it into Iceberg data lake and Postgres table(Destination)
  - **Technologies:** Python, PySpark, Iceberg, Nessie, Postgres, Shell Script, Great Expectation(GE)

- Product's Customer Reviews Data Batch Processing ELT Pipeline (Professional)
  - Streamlined data ingestion from Kafka into a data lake, enforced data quality checks using GE in Airflow, and orchestrated a Spark batch pipeline to implement transformations and aggregations, ensuring clean data is delivered to a Postgres database with 100% assurance that the data has successfully reached the Postgres DB.
  - **Technologies:** PySpark, Great Expectation(GE), Iceberg, Airflow, LLM, NLP, Python, Postgres

- Custom Data Quality Rules in GE with PySpark (Professional)
  - When it comes to the stream pipeline, the QA is completed by the time of streaming. While working on the pipeline with Spark, certain fields have specific rules and conditions. In this case, we customise the Great Expectations (GE) tools by modifying the source code and implementing custom rules, which makes the end-to-end data processing 72% faster.
  - **Technologies:** PySpark, Great Expectation(GE), Python

- Pipeline Managing Alert System (Professional)
  - To monitor the pipeline and alert the team when there are discrepancies, we will implement a complete Python-based project that can be configured using a YAML file. This file will be included in the project repository. The system will scan the project to locate the YAML file and monitor the pipeline using the configurations specified within it. This approach helps to reduce pipeline downtime by 83% and improve availability by 67%.
  - **Technologies:** Python, Google Chat, API, Crone

Education

SHM Engineering College, Kadakkal — March 2016 - Jun 2020
Computer Science and Engineering — *CGPA: 6.83*

Classic ITI, Anchal — April 2014 - May 2015
Computer Operation & Programming Assistant — *MARKS: 72%*

Higher Secondary Education Kerala — March 2014
Maintenance And Repairs Of Domestic Appliances — *MARKS: 53%*