NOORUL HUDA KHAN

📞 +919990734395 ✉ noorulhuda101099@gmail.com in linkedin.com/noorul ⌨ https://github.com/THENHKHAN

Experience

Fluree — Associate Data Engineer (Full-time)   **Aug 2023 – present**
Data Engineer   *Bangalore, India*

- Added functionalities and tools to convert TTL, RDF, and OWL files into organized formats, enhancing overall data management efficiency.
- Enhanced concurrent data processing in a Spark-based application by implementing logically isolated SparkSession instances for each actor using Akka. Applied session isolation techniques, ensuring seamless parallel execution of multiple actors, which improved scalability, performance, and resource efficiency.
- Developed a Python-based tool to convert Microsoft Dynamics 365 Common Data Model (CDM) JSON schemas into JSON-LD format compatible with FlureeDB, enabling semantic data modeling and easier graph database integration.
- Automated batch processing of entity definitions, including schema versioning, RDF class/property generation, and structured ontology linking using CDM standards. Aligned the schema transformation process with FlureeDB's data modeling requirements to support supply chain and inventory data structures in a graph-based context.
- Engineered a robust address normalization pipeline using LibPostal, pandas, and pycountry, introducing Unicode-safe cleaning, duplicate removal, and fallback logic for missing countries, ensuring accuracy and resilience across `1M+` records.
- Optimized runtime by **60%** by caching country lookups with `lru_cache`, replacing slow loops with vectorized operations, and integrating parallel-safe address cleaning via `swifter`.
- I built a PySpark pipeline that uses `SentenceTransformers` to batch-encode millions of records, stores them in a `FAISS` index, and performs fast nearest-neighbor searches for deduplication. I `optimized` it with batch processing, distance thresholds, and Spark SQL post-processing to handle `large datasets` in minutes instead of hours.

Fluree — Data Engineering Intern   **Feb 2023 – Aug 2023**
Internship   *Bangalore, India*

- Designed and implemented address parsers in Julia and C, optimizing data extraction and contributing to improved geospatial processing within the system.
- Engineered custom queries, algorithms, and APIs for an internal JSON-LD database using Scala, Spark, and Apache Jena, optimizing data retrieval processes. Engineered custom prompts and trained ChatGPT to understand and execute FlureeDB queries, improving efficiency and user experience.

Technical Skills

Languages: Python, Scala, Spark, Julia, PySpark, Sparql, Java, Javascript
Backend: Spring Boot, Fast-API, Familiar with (Flask, Django), SQL, PostgreSQL, MongoDB
Technologies: Docker, Databricks, AWS, Azure, Linux, Git, Github, Spring Boot, GraphQL, Postman
AI Models: all-MiniLM-L6-v2 (To Generate Embeddings), local LLaMA model (Q & A), mistral-7b-instruct-v0.1.Q4_K_M Hugging Face model, MCP-Server

Education

Aligarh Muslim University   **June. 2021 – July 2023**
Masters of Computer Applications (MCA) : 8.4 CGPA   *AMU, India*
Jamia Millia Islamia   **June. 2017 – July 2020**
B.sc(Hons) Physics : 8.7 CGPA   *JMI, India*

Projects

Build Real-Time REST APIs with Spring Boot -Blog App | *Java, Hibernate, PostgreSQL, Docker*   ⌨ **Github**

- Build Rest API's for Blog App Using Spring Boot, Spring Security, JWT , Spring Data JPA (Hibernate), PostgreSQL Database.
- Implemented Pagination and Sorting, Login and Sign-up, REST API Validation, Handled exceptions globally as well as specific.

Task Reminder Application | *fastapi, oauth2, react.js, mongodb, javascript*   ⌨ **Github**

- Developed a Task Reminder Application with FastAPI, OAuth2, and MongoDB, demonstrating proficiency in backend development.
- Implemented a responsive front end using React.js, enhancing user experience and interaction in the Task Reminder Application.