Avinash Madhukar Pawar

Bloomington, Indiana, USA
+1 (812)272-0824 | mail.avinashpawar@gmail.com | LinkedIn: avinashmpawar | Github: git-avinashpawar | Portfolio: Avinashpawar.dev

EDUCATION:

Master of Science in Data Science — **August 2021 – May 2023**
Indiana University, Bloomington — Indiana, USA
Coursework: Statistics, Machine Learning, Cloud Computing, Advanced Database Concepts, High-Performance Computing, Bioengineering.

Bachelor of Technology in Computer Science — **June 2016 – March 2020**
Shivaji University, Kolhapur — Kolhapur, India
Coursework: Distributed Systems, Operating System, Computer Networking, Database Management Systems, Algorithms, Microprocessors.

SKILLS:

Languages:	Python, SQL, C++ and JavaScript.
Databases:	MySQL, PostgreSQL, Hadoop, Spark, and BigQuery.
Visualization Tools:	Tableau, plotly, ggplot, Matplotlib, Seaborn, and PowerBI.
ML Algorithms:	Regression, Classification, Clustering, Decision Trees, and Neural Networks.
Cloud:	Linux, AWS(S3, EC2, Lambda), Cloud native technologies, Docker, and Kubernetes.
Miscellaneous:	Informatica, Snowflake, Agile, and Anaconda.

EXPERIENCE:

Research Data Analyst — **October 2023 – June 2024**
Indiana University - Kelley School of Business , Bloomington — Indiana, USA
- Digitized invoice data from PDFs into a centralized database using SQL and Python, generating financial insights and reducing processing time by 30% with automated scripts.
- Enhanced data accuracy and integrity through robust validation and cleansing, and presented analysis results via **Tableau** and **Excel**.
- Collaborated with cross-functional teams on **data scraping**, improved database architecture, and developed standardized data templates.

Data Analyst (Metadata Content Analyst) — **October 2021 – May 2023**
Indiana University, Bloomington — Indiana, USA
- Leveraged **SQL** to generate detailed reports for metadata discrepancies, assisting in the identification of unclean or incomplete records.
- Enhanced metadata quality by performing **ETL** processes, involving data cleansing and preprocessing, utilizing **Python** and **Excel**.
- Achieved a remarkable **40%** improvement in metadata quality by implementing systematic data cleansing techniques.
- Expedited metadata processing by **30%** through streamlining data preprocessing workflows and optimizing Excel functions.
- Aided in the development of standardized **metadata** templates, leading to consistency across the library catalog and improving user experience.

Software Engineer — **May 2019 - August 2021**
Digital Microsys Technologies, Kolhapur — Kolhapur, India
- Designed and maintained **scalable database solutions** for mission-critical applications, ensuring optimal performance and high availability.
- Optimized SQL scripts resulting in a **20%** reduction in query execution time and a **12%** refinement in overall database performance.
- Integrated **RESTful** API web services for precise data retrieval and storage, optimizing external data source interactions.
- Collaborated on developing web applications for a local grocery store and a hotel inventory management system using Django and MySQL. Implemented seamless **e-commerce features** including payment gateway integration, order tracking, and inventory management.
- Architected a data pipeline using **Python** and **Selenium** to automate data scraping, preprocessing, and modeling of utility data.

PROJECTS:

COVID-19 Lexicon in Media: An Analytical Perspective | Github
- Directed and led a cross-functional team in the creation of a dynamic **dashboard**, visualizing and analyzing extensive COVID-19 media data.
- Utilized ETL techniques and sophisticated **data pipelines** to process and integrate data from the GDELT dataset, totalling **700 GB**.
- Provided stakeholders an information-rich dashboard to uncover media trends and **patterns**, showcasing expertise in data integration, visualization, and analysis for nuanced pandemic insight. Utilized **GCP**, **BigQuery**, and **Tableau** to translate raw data into valuable insights.

Parallel K-means Accelerator for multidimensional data | Github
- Architected **K-Means Accelerator:** a high-performance parallel K-means clustering solution for multidimensional data using C++.
- Achieved dramatic speedups for K-means clustering of high-dimensional datasets by harnessing efficient multithreaded (**OpenMP**) and distributed-memory (**MPI**) parallelization on a supercomputer.
- Scaled the solution to a massive 140-node 64-core **supercomputer**, enabling ultrafast processing of colossal, multidimensional datasets.
- Slashed K-means clustering computation time, facilitating potential large-scale deployments on **1000-node** supercomputers.

Distributed Textbook Search Engine: MapReduce, Cloud Integration, and ETL Pipelines | Github
- Engineered a sophisticated **MapReduce**-based search engine for over **1000** textbooks, integrating ETL pipelines for data acquisition.
- Applied **GCP**, **Node.js** and Google **Cloud Functions** to deploy Mapper and Reducer components, optimizing scalability.
- Built an innovative web interface featuring rapid **sub-second search** results and advanced batch search via file links, streamlining efficiency.
- Showcased versatility in merging cloud deployment, ETL architecture, user-centric interface design, distributed computing, and data engineering.

ACHIEVEMENTS & CERTIFICATIONS:

Secretary | Data Science Club at IU — October 2021 – May 2023
Google Advanced Data Analytics Professional Certificate | Google | Certificate Link — June 2023
Winner**, AWS Game Day challenge** | AWS | Indiana Statewide IT Conference — April 2023