Avinash Madhukar Pawar

Indianapolis, Indiana, USA
Mobile: +1 (812)272-0824 | Email: 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, Signal Processing, Bioengineering.*
Achievements: *Secretary, Data Science Club | Google Advanced Data Analytics Professional Certificate, Link | Winner, AWS Game Day challenge, Link*

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, Data Mining, Algorithms, Microprocessors.

SKILLS:

Programming Languages	:	Python, SQL, C++ and JavaScript, R, Shell Scripting.
Fundamentals	:	Data modelling, Data quality, Query Optimization, Automation, Custom ETL, CI/CD, Data Warehousing.
Databases	:	MySQL, PostgreSQL, Hadoop, Spark, BigQuery MongoDB, Firebase, Google Cloud Storage, DynamoDB.
Visualization Tools	:	Tableau, plotly, ggplot, Matplotlib, Seaborn, PowerBI, Excel.
Machine Learning Tools	:	SciPy, Scikit, Pandas, NumPy, PyTorch, Regression, Classification, Clustering, Decision Trees, Neural Networks.
Cloud Technologies	:	Linux/Unix, AWS (S3, EC2, Lambda), Google Cloud Platform, Cloud native technologies, Docker, Kubernetes.
Generative AI	:	LangChain, Transformers (Hugging Face), OpenAI API, Google Gemini API, Streamlit.
Miscellaneous	:	Informatica, Git, Apache Spark, Apache Kafka, Apache Tomcat, Snowflake, Hadoop, MapReduce, Hive, Yarn.

EXPERIENCE:

Data Engineer | CVS Health | Bloomington, Indiana, USA **October 2023 – Present**

- Led the development of scalable data pipelines and optimized custom ETL processes using **Python** and **Apache Spark**, increasing data processing efficiency by **60%**. Automated CI/CD pipelines to handle up to **4TB of data daily** from diverse sources.
- Migrated legacy data systems to **Snowflake**, managing a secure data lake infrastructure that cut query times by **40%** and ensuring data security with robust quality checks using **Apache Airflow**. Performed performance tuning and query optimization to enhance database efficiency.
- Authored and updated comprehensive documentation for ETL processes, data pipelines, and system architecture; facilitated seamless handovers while enhancing team collaboration that contributed to a **30% reduction** in onboarding time for new engineers.
- Integrated five new data sources into the data ecosystem, providing **insights that led to a 15% increase in actionable recommendations**.

Data Engineer | Indiana University - Kelley School of Business | Bloomington, Indiana, USA **October 2021 – May 2023**

- **Automated** the digitization of invoice data from PDFs into a centralized database using **SQL** and **Python**, **reducing processing time by 30%**.
- Enhanced data accuracy and integrity via robust validation and cleansing; presented analysis results via **Tableau** and **Excel**.
- Collaborated with cross-functional teams to redesign database architecture, develop data templates, and improve data scraping methodologies.
- **Executed SQL queries** to extract and analyze **1M+ financial transactions** from multiple tables. **Collaborated with stakeholders** to identify data sources, improving accuracy by **20%**. Utilized **Excel** (Pivot Tables, VLOOKUP, Macros) to ensure data compliance and integrity.
- Developed interactive **Tableau dashboards** with calculated fields and KPIs, reducing manual data analysis efforts by **30%**.

Software Engineer, Data Platform | Tata Consultancy Services | Pune, India **May 2019 - August 2021**

- Designed and maintained scalable database solutions for mission-critical applications, ensuring **high availability** and optimal performance.
- Optimized SQL queries, achieving a **20% reduction** in query execution time and improving overall database performance by **12%**.
- 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 Python-based data pipeline using **Selenium** to automate data scraping, preprocessing, and modeling of utility data.

PROJECTS:

PragyaYantra: A Generative AI web application | Github | Website

- Launched a sophisticated AI web application showcasing a range of **cutting-edge AI capabilities**, powered by the Google Gemini API. This application provides users with intuitive tools for text generation, intelligent dialogues, file handling, and more.
- Utilized HTML, JavaScript, CSS and **NodeJS** to create an intuitive user interface for seamless interaction across multiple AI modules.
- Integrated the **Google Gemini API** to power real-time text generation, conversation simulation, document analysis, and code generation.
- Focused on user experience and responsiveness, featuring interactive elements to enhance engagement and streamline content generation.

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 256-node 64-core **supercomputer**, enabling ultrafast processing of colossal, multidimensional datasets.
- Slashed K-means clustering computation time, facilitating potential large-scale deployments on more than **1000-node** supercomputers.

Distributed 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.