Yash Rai

☎ 9026332517 ✉ Yashrai1224@gmail.com linkedin.com/in/yashrai1224 github.com/Yash-rai-29 portfolio

Summary

— Data Engineer specializing in scalable data pipelines and big data technologies. Skilled in optimizing data workflows, enhancing system performance, and reducing costs. Proficient in using advanced tools to manage and process large datasets. Seeking to apply technical expertise and innovative problem-solving skills to a dynamic organization.

Technical Proficiencies

Programming Languages: Python, C++, C
Development Tools and Frameworks: Git, GitHub, Jira, Visual Studio Code, Docker, Google Kubernetes Engine (GKE)
Big Data and Cloud: GCP, Azure, ADF, DataBricks, SQL, Apache Beam, BigQuery, Bigtable, Cloud Function, Dataflow, Airflow, Pub/Sub, Dataform

Professional Experience

Clarity **Feb 2024 – Jun 2024**
Software Development Engineer Intern

- Developed a **Customer Data Platform (CDP)** web application using **React.js** and **Node.js**, significantly enhancing user data management. The platform fetches event data from **Bigtable** based on user ID, lists all events, and provides filters for specific time ranges to review user payment or event history.
- Independently implemented a **real-time event pipeline ETL** using **Python** and **Apache Beam**, reading over **5 million user events** daily from **Pub/Sub**, transforming the data, and loading it into **BigQuery** and **Bigtable**. This solution improved data processing efficiency and reduced data handling costs by **10%**.
- Read raw data from **BigQuery** and transformed it using **SQLX** in **Google Dataform**. Implemented an **ETL** process to convert raw data into incremental tables, enhancing data accessibility and reducing query times.

BinPlus Technologies **Nov 2023 – Jan 2024**
UI/UX Developer Intern

- Enhanced a **casino game project** by integrating **Socket.IO** components, enabling real-time gameplay connectivity and improving user engagement.
- Architected and implemented a **betting website** from the ground up for a client, integrating multiple APIs for dynamic user management and real-time data updates.
- Designed and developed the frontend using **React.js** and **Bootstrap**, ensuring a fully responsive and dynamic user interface across all devices.
- Implemented robust **authentication and authorization** mechanisms using **Node.js**, securing APIs with token-based authentication to protect user data.

Abhyaz **Dec 2022 – May 2023**
Web Developer Intern

- Developed and maintained **three dynamic and responsive websites** on the Zoho platform, utilizing Zoho Sites to create user-friendly web applications.
- Implemented various features and functionalities to enhance user experience, including **interactive forms**, **calendar event markers**, and other dynamic elements.
- Collaborated with a team to develop a fully functional website for **Matlab Technology** from scratch in just **2 weeks** using Zoho Site Editor, ensuring timely delivery and high-quality output.

Education

SR Group of Institutions, Jhansi **2020 – 2024**
B. Tech in Computer Science & Engineering *AKTU University*

Projects

Real-Time Streaming Data Pipeline
Python, Apache Beam, Google Cloud Pub/Sub, Apache Kafka, BigQuery, Bigtable, Dataflow

- Developed a **real-time streaming pipeline** to read data from **Google Pub/Sub** topics sourced from **Apache Kafka**.
- Used **Apache Beam** (Python) to transform and clean data, ensuring proper **partitioning** and **schema management**.
- Wrote cleaned data into **BigQuery** with proper **partitioning** and **schema management**.
- Stored specific topic data in **Bigtable** for **high-throughput** and **low-latency** access.
- Deployed and managed the pipeline on **Google Dataflow** for real-time processing and monitoring.

Raw Data to Incremental Table
Google Dataform, BigQuery

- Implemented a **ETL** using **Google Dataform** to transform and clean raw data from various raw data table in Bigquery.
- Transformed data and created views, then used these views to build **incremental tables** for analytics.
- Ensured **data integrity** using **SQLX**, applying various joins and row functions to validate and transform data.
- Automated the ETL using **CronJob** for scheduled runs, with version control managed via **Git**.