Abhay Kumar

Bengaluru, India | kumarabhay.de@gmail.com | +917019751217

linkedin.com/in/abhay-kumar-b3561b144 | github.com/abhay-kum

Summary

A highly detail-oriented and **results-driven Data Engineer with 6+ years** of expertise in **architecting and delivering robust** big data solutions for **FinTech and E-commerce**. My experience spans building **scalable real-time and batch data pipelines with Spark, Kafka, NiFi, and Airflow on the Hadoop ecosystem**, leveraging AWS cloud services like **S3, EMR, Glue, and Lambda**. I excel at handling diverse data sources for **analytics, reporting, and ML initiatives, consistently driving business insights** and leading successful projects.

Technical Skills

- **Languages & Scripting:** Python, Bash, SQL
- **Big Data:** PySpark, Hive, Impala, Apache Iceberg, HDFS, Hadoop, Sqoop
- **Data Tools:** Apache NiFi, Airflow, Dreamio, DBT
- **Cloud & Infra:** AWS, GCP, Azure, Docker
- **Web Scraping:** Scrapy, BeautifulSoup
- **Databases:** MongoDB, Oracle, ElasticSearch
- **CICD/Workflow Tools/Version Control:** Jenkins, Control-M, Crontab, Github
- **GEN AI:** LLM (OpenAI, Gemini), RAG, Vector Database, Embedding

Experience

Data Engineer-2, IDFC FIRST BANK – Bengaluru, India Mar 2022 – Present

- Architected and implemented a scalable, automated data ingestion framework capable of processing **1000+ tables**, facilitating **multi-stage data processing across staging, refined, and raw layers** using **Airflow for orchestration and Sqoop, Spark, and Hive scripts** for efficient data transfer and transformation.
- Designed and implemented an automated, **event-driven framework for internal-audit reporting**, generating **75+ branch-level reports**. This initiative reduced **Turnaround Time (TAT) from 7 days to 1 day** and **eliminated manual intervention**, significantly improving efficiency.
- Created **40+ user-friendly Impala scripts** with variable inputs, **enabling non-technical stakeholders to efficiently retrieve critical LEA/MHA reports without direct technical assistance**.
- Developed and automated pipelines for extracting **beneficiary information from multi-source account statements**, streamlining data preparation for LEA/MHA reporting and **reducing TAT by over 60% (from 1 day to 3-4 hours)**.
- Developed pipelines that efficiently collected crucial **customer portfolio and onboarding data**, including **account numbers, loan distributions, NPA details, app scores, demographics, and transaction amounts, for RBI's required monthly and quarterly reports.**
- **Tech Stack: Python, Pyspark, SQL, Apache-Airflow, Apahce-Nifi, Sqoop, Hive, impala, AWS, Jenkins, Git.**

Data Engineer, TextMercato – Bengaluru, India Nov 2020 – Feb 2022

- **Engineered & deployed a scalable data pipeline for efficient collection and processing of unstructured data** (PDFs, images, Excel, websites).
- **Leveraged advanced AI/ML techniques**, including **OCR (PaddlePaddle) and OpenAI-based paraphrasing models**, for data transformation.
- Generated comprehensive **product catalogs for major e-commerce platforms (Amazon, Flipkart, Myntra)**.
- Developed **robust & scalable APIs** using FastAPI to support **data ingestion, processing, and retrieval**.
- **Tech Stack: Python, Scrapy, Azure, FastAPI, OCR (PaddlePaddle), OpenAI API, MongoDB.**

Data Engineer, Greendeck – Indore, India Mar 2019 – Oct 2020

- Built **competitor intelligence systems** for e-commerce clients.
- Built a robust system for continuous, efficient data acquisition and storage through scraping more than **100+ e-commerce sites**.
- Leveraged data **snapshots** and promotional **newsletters to** continuously monitor market pricing and promotional strategies..
- Used **Dockerized microservices with FastAPI deployed on GCP.**
- **Tech Stack: Python, GCP, Microservices, Docker, ELK, Scrapy, FastAPI, MongoDB, Redis.**

Education

PG Diploma in Data Science, IIIT Bengaluru Jan 2022 – Feb 2023

- **CGPA**: 3.49/4.0
- **Coursework:** Machine Learning, Data Engineering

B.Tech in Computer Science Engineering, IIIT Dharwad Aug 2015 – Apr 2019

- **CGPA**: 8.2/10.0
- **Coursework:** Operating System, DSA, Software Engineering, Data Science

Certifications

The Complete DBT (Data Build Tool) Bootcamp: Zero to Hero	Udemy	June, 2025
Introduction to Data Science in Python	Coursera	June, 2018
Applied Plotting, Charting & Data Representation in Python	Coursera	June, 2018

Awards

- Award of Excellence in Data Platform Migration for leading the successful Cloudera migration at IDFC FIRST Bank