

Sahil Pandita

Lead Data Engineer – Big Data Platforms, ML Data Pipelines & AI-Powered Data Tools

pandita.sahil1@gmail.com · sahilpandita.com · linkedin.com/in/sahil-pandita-90994692 · github.com/Sahilpandita1

I'm a data engineer who has spent 12 years turning messy, high-volume telecom and enterprise data into reliable, governed systems that are ready for analytics and ML. I've led migrations from Oracle to Hadoop data lakes and built CDC pipelines from GoldenGate and Solace queues into Hive. Today I lead the DARTS BI reporting team at Airtel X Labs and work on the data platform behind Airtel's AI-powered spam detection.

12+

years in data engineering

300M+

telecom users protected
by the spam platform I
work on

600M+ × ~70

records × features
monitored for drift per
run

202/202

automated tests passing
on my LLM data agent

4

companies: Airtel,
Capgemini, Exusia,
Accenture

EXPERIENCE

Lead Engineer, Airtel X Labs

Dec 2021 – Present · 4 yrs 11 mos

Airtel Digital India Pvt. Ltd. · India

- Lead engineer for DARTS BI Reporting, managing and mentoring a team that delivers business-critical analytics.
- Played a key role in India's first AI-powered spam detection solution, protecting 300M+ telecom users from spam and fraud; contributed to system architecture and integration with telecom networks.
- Designed and built a scalable, reusable PySpark framework that detects and monitors data drift across multiple ML pipelines and raises early alerts when data changes.
- Built an AI-powered natural-language data agent for the spam-detection team.
- Led GCP Reporting pipelines in Ab Initio to extract, transform and reintegrate reports into the GCP ecosystem.
- Delivered SMS Insights, bringing transparency to the bulk SMS aggregator market, with automated CSV reports pushed to AWS S3.
- Built a continuous CDC pipeline for Siebel postpaid and Telemedia reporting, ingesting from Solace queues and Oracle GoldenGate into Hive for daily reports.
- Developed the AIO Reporting framework to standardise and automate enterprise reporting.
- Implemented PII data encryption across the DARTS ecosystem.

PySpark · Spark on YARN · Hive · HDFS · Alluxio · Ab Initio · Aerospike · Airflow · GoldenGate CDC · Solace · GCP · AWS S3 · Python · LLMs

Software Development Consultant, Capgemini India

Aug 2018 – Dec 2021 · 3 yrs 5 mos

Pune

- Led a data lake migration, extracting data from Oracle source systems into Hadoop-based data lakes.
- Built utilities to automate DML creation and test-data generation, cutting manual effort and errors.
- Designed batch processes that analysed every migrated table and produced validation reports.
- Implemented Masspay ANP and CDC pipelines for reliable incremental data processing.

Ab Initio · Hadoop · Hive · Oracle · CDC · Unix shell

Senior Analyst, Exusia

Jul 2017 – Aug 2018 · 1 yr 2 mos

Pune

- Executed the PAMS data migration from Oracle (via stored procedures) into Netezza (COGS).
- Built parameter sets (psets) with Express>It inside an in-house Ab Initio framework.

Ab Initio · Express>It · Oracle · Netezza · PL/SQL

Accenture

Gurugram

Aug 2014 – Jun 2017 · 2 yrs 11 mos

Application Development Analyst Dec 2015 – Jun 2017 · 1 yr 7 mos

Associate Software Engineer Aug 2014 – Dec 2015 · 1 yr 5 mos

- Worked on Global Data Quality programmes to analyse, measure and report on critical data elements; authored data quality rules.
- Migrated large-scale DQ and EVM projects from AIX to Linux with zero data loss.
- Built an Ab Initio regression utility comparing AIX and Linux outputs and table loads, generating Excel migration reports.
- Received the Accenture Excellence Award (2016).

Ab Initio · Data Quality · AIX · Linux · Unix shell

SELECTED PROJECTS

AI-Powered Natural-Language Data Agent

AI / LLM

A conversational agent that turns plain-English questions into Hive SQL, runs them on the cluster and replies with results, charts or files. **202/202** automated tests passing; **35/44** pass rate when the harness started.

Python · PySpark · Hive · HDFS · YARN · LLMs · Gemma · Prompt Engineering · Telegram API · Confluence

Data Drift Monitoring Framework for ML Pipelines

ML Data Platform

A reusable PySpark framework that compares feature distributions between time windows and raises early alerts, across multiple spam ML pipelines. **600M+** records per run; **~70** features tested for drift.

PySpark · Spark on YARN · Statistics · MLOps · Parquet · HDFS

Spam Detection Data Platform

ML Data Platform

The data platform behind India's first AI-powered spam detection solution, protecting 300M+ Airtel users from spam and fraud calls and SMS. **1st** AI-powered spam detection in India.

PySpark · Hive · Aerospike · Airflow · YARN · Ab Initio · ML features

More case studies at sahilpandita.com:

- Siebel CDC Reporting Pipeline
- GCP Reporting & SMS Insights
- Oracle Data Lake Migration
- Oracle Hierarchical Logic → PySpark
- Ab Initio Attribute Onboarding Automation
- Self-Healing Metadata Watchdog for Hive/Alluxio
- Enterprise PII Encryption & Reporting Standardisation
- AIX → Linux Migration Regression Utility
- Local LLM Agent

SKILLS

Big Data & Processing

PySpark, Apache Spark, Spark on YARN, Hive, HDFS, Alluxio, Hadoop

ETL & Data Integration

Ab Initio GDE, Express>It, EME, CDC (Oracle GoldenGate), Solace Queue

Databases & Storage

Oracle, Aerospike, Netezza, Parquet

Orchestration & Ops

Apache Airflow, Unix/Linux shell scripting, cron

Cloud

GCP (reporting), AWS S3

Languages

SQL, Python, PL/SQL

ML & AI Enablement

Feature-ready datasets, Data drift monitoring (KL, PSI, KS), LLM agents, NL-to-SQL, Prompt engineering, Test harnesses for LLM apps

Data Management

Data quality, Data governance, PII encryption, Data migration & reconciliation, Performance tuning

Leadership

Team leadership & mentoring, Stakeholder reporting

EDUCATION

Post Graduate Program in AI & Machine Learning

2023–2024

Great Lakes Executive Learning & The University of Texas at Austin · CGPA 3.79/4.0

B.E., Information Technology

2010–2014

Sinhgad Academy of Engineering, University of Pune