SAI SAMPATH BODDU

sampathusa1333@gmail.com | +1 (312) 869-4651

SUMMARY

- Data Engineer with three years blending ETL & ELT craftsmanship and pragmatic machine-learning enablement, delivering dependable data pipelines, feature-ready tables, and stakeholder-friendly dashboards across healthcare and logistics domains.
- Skilled at translating messy operational extracts into analytics-grade datasets, then augmenting those assets with scikit-learn, Hugging Face models that score risk, churn, or text sentiment directly inside scheduled workflows.
- Comfortable orchestrating batch jobs through Airflow, Prefect, storing assets in cloud-agnostic object storage, and surfacing trusted metrics in Tableau, Power BI, or Metabase reports.
- Known for establishing automated data-quality gates, model-performance monitors, and Slack or Teams alerting channels that keep business owners confident in both raw numbers and the predictive signals layered upon them.
- Communicates complex data or model behaviour in plain language, pairing Jupyter prototypes with Lucidchart lineage diagrams so non-technical partners quickly grasp how sources become actionable, governed insights.

TECHNICAL SKILLS

Languages & Query: Python (Pandas, NumPy, requests), SQL (PostgreSQL, SQL Server, BigQuery), Bash
Data Integration: Airflow DAGs, Prefect flows, Databricks Jobs, Python scripts
Databases & Storage: PostgreSQL, MySQL, SQL Server, BigQuery, Parquet/CSV
Transformation Libraries: Pandas, pyarrow, dbt (core), openpyxl, csvkit
Business Intelligence: Tableau, Power BI, Metabase, matplotlib
Scheduling & Orchestration: Airflow (self-hosted & managed), Prefect Cloud, Databricks Workflows
Cloud Platforms: Azure SQL DB, on-prem VMware clusters, AWS, Google Cloud Storage, BigQuery
DevOps Basics: Git/GitHub, GitHub Actions, Azure Pipelines, Docker
Data Quality & Testing: Great Expectations, dbt tests, SQL sanity checks
Documentation & Diagrams: Confluence, Markdown READMEs, Lucidchart ERDs, Mermaid flows
Excel Power-User: Pivot tables, Power Query, XLOOKUP, VBA macros
Versioned Notebooks: Databricks Repos, Jupyter, nbconvert
Monitoring & Alerts: Grafana, Prometheus, Datadog, Slack webhooks
Security & Governance: Row-level security in SQL, column masking, GDPR tagging
Collaboration Tools: JIRA, Slack, MS Teams
Publications: “Agile Data Science and its Relevance”, IRJMETS, 2021

WORK EXPERIENCE

Data Engineer, Client: ***Optum Inc*** Jan 2024 – Present

- Orchestrated nightly Airflow DAGs ingesting multi-gigabyte claim extracts from secure SFTP, validating file structures, and writing partitioned Parquet snapshots to cloud object storage for downstream analytics consumption.
- Authored advanced SQL-based transformations employing nested Common Table Expressions and window functions to compute member cost ratios, readmission frequencies, and denial rate benchmarks across diverse service lines.
- Published interactive Tableau dashboards where finance stakeholders explore five-year trend lines, drill into provider group performance, and export filtered views directly to PowerPoint decks for quarterly reviews.
- Replaced a manual Excel refresh routine with Power BI dataflows that automatically aggregate seven departmental spreadsheets, reducing repetitive analyst labour by roughly six hours each operational week.
- Embedded Great Expectations tests inside ingestion DAGs, enforcing column-type checks, regex policy-ID validation, and aggregate row-count reconciliation before declaring data assets production-ready.
- Connected Prometheus exporters to Airflow task instances, visualising runtime, success counts, and SLA breaches in Grafana, then dispatching Slack alerts whenever predefined latency thresholds were exceeded.
- Implemented tiered lifecycle policies on the object store, migrating raw landing-zone files to lower-cost archival tiers after three months, leading to approximately thirty-eight-percent overall storage savings.
- Collaborated with actuarial scientists to build a statsmodels ARIMA forecasting notebook predicting membership growth, logging back-test metrics with MLFlow, and exporting quarterly projections into controlled Excel templates.
- Containerised auxiliary data-cleaning scripts using Docker-Compose, ensuring developers reproduce identical environments locally and avoiding “works on my machine” discrepancies during peer-review sessions.
- Authored a comprehensive Confluence data dictionary covering forty curated tables, each documenting column definitions, refresh cadence, primary keys, and example analytic queries for self-service clarity.
- Acted as JIRA sprint reporter, updating burndown charts, recalibrating story points post-stand-up, and communicating impediments to the broader data leadership triad.

- Implemented row-level security on a managed SQL warehouse, ensuring Protected Health Information columns surface only to clinicians with explicit approval, thereby passing internal compliance audits.
- Wrote a scheduled Python script posting nightly pipeline statistics—rows processed, error counts, model-prediction summary AUC—to a Slack channel consumed by operations stakeholders each morning.
- Developed an Excel VBA macro that connects via ODBC, fetches aggregated financial summaries, and populates board-ready worksheets in under thirty seconds, replacing previously fragile copy-paste workflows.
- Presented a business review deck combining Tableau screenshots and Loom-recorded walkthroughs, receiving executive praise for translating technical pipeline details into accessible financial impact narratives.
- Mentored two summer interns through Git branching etiquette, SQL anti-patterns, and basic scikit-learn usage, culminating in each intern shipping a small production report enhancement before program completion.

Data Engineer, Client: ***Grepthor Software Solutions Pvt Ltd*** May 2021 – Jul 2022

- Consolidated twelve legacy CSV feeds into PostgreSQL staging schemas using Python loaders that automatically handled delimiter ambiguities, character encoding mismatches, and inconsistent quote escapes across vendor files.
- Designed an Airflow DAG comprising extraction, transformation, and load tasks that join shipment, inventory, and billing events into a consolidated fact table, meeting a forty-five-minute end-to-end service-level objective.
- Created a reusable Pandas data-cleaning toolkit supporting null imputation, ISO-8601 date standardisation, duplicate suppression, and categorical code mapping, now leveraged by four separate client engagements.
- Built Metabase dashboards visualising on-time delivery percentage, average miles per stop, and warehouse dwell time, enabling account managers to negotiate service-level improvements with transportation partners.
- Added strategic B-tree and partial indexes plus routine VACUUM and ANALYZE maintenance jobs on high-volume tables, reducing typical query runtimes by approximately sixty percent.
- Produced Lucidchart ERDs illustrating raw ingestion layers, operational data stores, and analytic presentation schemas, accelerating onboarding for new developers unfamiliar with logistics domain nuances.
- Implemented a lightweight filesystem crawler that registers inbound JSON parcels into a central metadata catalog, enabling ad-hoc SQL exploration through an embedded SQLPad interface.
- Wrote a Python REST client that retrieves carrier tracking events every thirty minutes, appends them into incremental load tables, and notifies the operations channel upon status code anomalies.
- Configured a Grafana dashboard monitoring job runtime percentile distributions, failure counts, and DAG success ratios, providing management with real-time visibility into data-pipeline health.
- Established GitHub Actions workflows executing flake8 lint, pytest suites, and schema-change diff tools, blocking pull-request merges whenever code fell below eighty-five-percent unit-test coverage.
- Re-implemented a complex Excel cost-allocation workbook using SQL window functions and CTEs, allowing analysts to run same-day variance scenarios without waiting for overnight recalculations.
- Standardised PostgreSQL column comments and dbt documentation blocks, empowering non-technical team members to self-serve definitions directly within Metabase’s built-in data-browser panels.
- Provisioned a Docker-Compose stack containing Postgres, pgAdmin, and a mock logistics API, streamlining local QA and demonstration environments used by product and sales teams.
- Negotiated with infrastructure administrators to provision a read-only logical replication slot, enabling analysts to query near-real-time production data without risking transactional contention.
- Facilitated fortnightly “Show and Tell” sessions where dashboard enhancements and model improvements were demonstrated, soliciting feedback and reinforcing an iterative delivery culture.
- Archived approximately three terabytes of obsolete application logs to cold object storage tiers, yielding projected annual savings close to four thousand US dollars.

Data Engineer, Client: ***Inductive Quotient Analytics India Pvt Ltd*** June 2020 – April 2021

- Developed Python ETL routines that parse HL7 laboratory result messages, apply regex-based field extractions, and store cleaned observations inside a secured PostgreSQL schema compliant with healthcare regulations.
- Authored SQL views that combine patient demographics, clinical visits, and insurance claims, powering daily operational dashboards consumed by nursing supervisors and case-management coordinators.
- Built an Excel pivot workbook enabling nurses to slice adverse-event counts by ward, shift, and attending physician, significantly decreasing ad-hoc reporting ticket volume.
- Executed schema-compare scripts before each deployment, identifying breaking column changes early in quality-assurance cycles and preventing downstream dashboard failures in production.
- Automated CSV-to-blob-storage ingestion followed by scheduled SQL warehouse loads using open-source rsync and cron jobs, later migrating schedules to Airflow for simplified dependency management.
- Implemented a row-count reconciliation process writing expected versus actual counts to a QA control table, surfacing three unnoticed data gaps within the first operational month.

- Assembled a Power BI report analysing medication adherence rates, revealing a twelve percent improvement following pharmacy process changes and prompting further intervention investment.
- Wrote Bash cron scripts to gzip and timestamp archive log files, freeing on-premises network-attached storage while retaining retrieval capability through symbolic-link manifests.
- Created an SQL warehouse user-defined function that masks protected-health-information fields, enabling analysts to perform exploratory work without violating patient-privacy guidelines.
- Configured pgAgent jobs executing routine ANALYZE statistics refresh and index rebuild tasks, keeping query planner estimates accurate as table volumes grew.
- Introduced Markdown README templates for every new repository, capturing purpose, inputs, outputs, schedules, and contact points, thereby streamlining hand-offs during holiday rotations.
- Drew Lucidchart lineage diagrams tracing raw interface tables through transformation layers into published marts, satisfying audit requirements and accelerating new-hire comprehension.
- Assisted senior engineers during a five-hundred-gigabyte migration from SQL Server to PostgreSQL using logical replication and validation checksums to ensure row-level accuracy.
- Filed weekly JIRA reports summarising defect resolutions, feature enhancements, and open blockers, consistently closing an average of seven issues each sprint.
- Presented analytical findings at monthly clinical round-table sessions, fielding questions regarding data freshness, filter logic, and interpretation caveats from medical directors.

EDUCATION

University of Illinois Chicago (UIC), Chicago, IL
MS in Management Information Systems
Indian Institute of Information Technology (IIIT), Sri City, India
Bachelor of Technology, Computer Science and Engineering