

Gaurav Gurjar

Senior Data Engineer

UAE Golden Visa holder — no sponsorship required
ggurjar333@gmail.com · linkedin.com/in/ggurjarsocl · GitHub @dataengineergaurav

SUMMARY

Senior Data Engineer and AI Data Platform Architect with 7+ years delivering cloud and governed AI data systems for insurance, public health, veterinary healthcare, cannabis analytics, capital markets, and government research. Strong in Python, SQL, AWS, Redshift, dbt, PySpark, Kafka, FastAPI, and Dagster; built 300+ production pipelines and contributed to a COVID-19 risk application used by 2M+ people.

INDUSTRIES

AI Governance Insurance Public Health Veterinary Healthcare Cannabis Analytics Capital Markets Government Research

PROFESSIONAL EXPERIENCE

Technical Advisor, AI Data Platform — Govern.ai (via Archetypal AI) 2026 → present

AI Governance · advisor

The AI data-platform function designs the governed data plane that turns policies and knowledge sources into structured, searchable artifacts for grounding and runtime enforcement, and builds the policy-aware routing, provenance, and governance-telemetry layers of a multi-provider browser agent.

- Designed governed AI data pipelines that convert policies and knowledge sources into structured, searchable artifacts for model grounding.
- Built policy-aware LLM routing, KB-only execution paths, provenance controls, and governance telemetry for a multi-provider browser agent.
- Developed automated testing, live-browser QA, and release workflows improving reliability across AI-provider integrations.

policy-pipelines LLM-routing grounding provenance governance-telemetry FastAPI pytest

Consultant, Senior Data Engineer — ISHIR (client: SageSure) 2024-04 → 2026-01

IT Services / Data & AI · employee

The SageSure engagement delivered insurance fact/dimension models, compliance datasets, and analytical marts for a catastrophe-exposed property managing general underwriter, building the dbt/Redshift data platform behind regulatory and business reporting.

- Delivered insurance fact/dimension models, compliance datasets, and analytical marts on Amazon Redshift using dbt and SQL.
- Implemented dbt data-quality tests and SQL integration tests; deployed daily/monthly incremental snapshots with scheduled Redshift automation.
- Developed reusable ETL packages, database modules, and typed FastAPI integrations.

dbt Redshift SQL AWS FastAPI pytest

Consultant, Data Engineering — Petfolk USA (concurrent with ISHIR) 2025-10 → 2026-01

Veterinary Healthcare · consultant

A modular, Dagster-based inventory-adjustment ETL package covering 46 clinic locations, validating location mappings and automating reproducible weekly inventory releases across the clinic-operations function.

- Authored a modular inventory-adjustment ETL package covering 46 clinic locations using Dagster, boto3, Pydantic, and API integrations.
- Validated location coverage and mappings, handled data exceptions, and automated reproducible weekly releases.

Dagster AWS Pydantic boto3 API-integrations

Data Engineer — CannaSpyGlass 2021-04 → 2024-03

Cannabis Market Data & Analytics · contractor

The data-engineering function built and operated production ingestion and transformation pipelines sourcing regulated cannabis businesses, geospatial data, APIs, and PDF datasets, producing analytics-ready outputs across AWS, GCP, and FastAPI services.

- Owned cloud ingestion and transformation for 300+ production pipelines spanning regulated cannabis businesses, geospatial data, APIs, and PDF sources.
- Produced analytics-ready datasets with Python, SQL, AWS Glue, PySpark, GCP, and FastAPI services.

AWS-Glue PySpark GCP FastAPI Python PDF-parsing geospatial

Data Scientist — AI Squared Inc. 2020-05 → 2021-03

Enterprise AI Platform · contractor

The data-science team built the statistical and data-science components for COVID19RiskTools.com, wrapping the iCARE R package in Python and powering AWS Lambda/RDS-backed APIs for a COVID-19 mortality-risk application.

- Researched and wrapped the iCARE R library; developed statistical, data-science, and backend components in R and Python.
- Built data pipelines and AWS Lambda/RDS-backed APIs for a public COVID-19 risk application.

R Python AWS-Lambda RDS iCARE

Data Engineer / Assistant Project Manager — Casepoint LLC 2018-08 → 2020-04

LegalTech / eDiscovery · employee

Two concurrent client engagements: (1) TradeTips - Glue/S3/Aurora data warehouse and real-time Kafka/FastAPI ingestion for a stock-signal platform, mentoring three junior engineers; (2) NHS UK Genetics - data-architecture blueprint and Kafka/NoSQL ingestion for a 2-20 PB precision-medicine infrastructure.

- Built Glue workflows, S3 storage pipelines, and Aurora data-warehouse components for stock signals and investment analytics (TradeTips).
- Developed real-time ingestion and scraping services with Python, Kafka, FastAPI, PySpark; guided three junior data engineers.
- Contributed a data-architecture blueprint and Python access tooling for a 2-20 PB NHS genetics infrastructure; built Kafka/NoSQL ingestion.

AWS-Glue Kafka FastAPI Aurora PySpark Python NoSQL

SELECTED CLIENT PROJECTS

SageSure Insurance Data Platform — SageSure · Insurance (Catastrophe-exposed Property)

Fact/dimension models, compliance datasets, and analytical marts on Amazon Redshift using dbt and SQL, with data-quality tests and incremental snapshots.

- Developed fact and dimension tables, compliance datasets, and analytical data marts on Amazon Redshift using dbt and SQL.
- Implemented dbt data-quality tests and SQL integration tests; deployed daily and monthly incremental snapshots.

SageSure ↗

RedshiftdbtSQL

CannaSpyGlass Cannabis Analytics Platform — CannaSpyGlass · Cannabis Market Data & Analytics

Cannabis market-data platform consolidating wholesalers, transporters, growers, dispensaries, Google Maps, and PDF datasets across the United States.

- Built and operated 300+ pipelines sourcing cannabis wholesalers, transporters, growers, dispensaries, Google Maps, and PDF datasets.
- Developed AWS Glue and PySpark ETL jobs, FastAPI-ready datasets, and GCP pipelines for essential-services location data across the US.

CannaSpyGlass ↗

AWS-GluePySparkGCPFastAPI

COVID-19 Mortality Risk Tool — COVID19RiskTools.com (Public Health) · Public Health

Statistical and data-science components in R/Python (wrapping the iCARE R package) with AWS Lambda/RDS-backed APIs for a public risk application.

- Developed statistical and data-science components in R and Python, wrapping iCARE R package functions in Python.
- Built AWS Lambda and RDS-backed data services for COVID19RiskTools.com, used by 2M+ people in the US.

COVID19RiskTools ↗

RPythonAWS-LambdaRDS

Petfolk USA Veterinary Inventory ETL — Petfolk USA · Veterinary Healthcare

Modular inventory-adjustment ETL with Dagster, Pydantic, and API integrations; validated mappings and weekly releases.

- Authored a modular inventory-adjustment ETL with Dagster, Pydantic, boto3, and API integrations.
- Validated location mappings and automated reproducible weekly releases.

Petfolk ↗

DagsterPydanticAWS

TradeTips Stock-Market Signals Platform — TradeTips · Capital Markets / Fintech

Glue workflows, S3 storage, and Aurora warehouse components with real-time Kafka/FastAPI ingestion and scraping services for stock signals.

- Built Glue workflows, S3 storage pipelines, and Aurora warehouse components for stock signals and investment analytics.
- Developed real-time ingestion and scraping services with Python, Kafka, FastAPI, PySpark.

TradeTips ↗

AWS-GlueKafkaFastAPIAurora

NHS UK Genetics Data Infrastructure — NHS UK (Genetics) · Public Health / Genomics

Data-architecture blueprint and Python access tooling for a genetics data infrastructure; Kafka/NoSQL ingestion from NHS source systems.

- Contributed to a data-architecture blueprint and Python access tooling for a genetics data infrastructure supporting precision medicine.
- Built Kafka and NoSQL ingestion and data-wrangling workflows from NHS source systems.

Genomics England ↗

KafkaNoSQLPython

SKILLS

AI Data Systems

- knowledge ingestion
- policy processing
- grounding
- retrieval
- LLM routing
- PII governance
- runtime telemetry

Data Engineering

- ETL/ELT
- dimensional modeling
- data quality
- incremental processing
- dbt Core
- Dagster
- Airflow
- Spark/PySpark
- Kafka

AWS

- S3
- Glue
- Lambda
- Redshift
- EMR
- Athena
- RDS/Aurora
- Lake Formation
- CloudWatch

Languages

- Python
- SQL
- R
- Bash
- Java
- TypeScript/JavaScript

Data Platforms

- PostgreSQL
- MySQL
- SQL Server
- DuckDB
- SQLite
- Polars
- Pandas
- Power BI
- Tableau

Engineering

- FastAPI
- Pydantic
- SQLAlchemy
- pytest
- Docker
- Kubernetes
- GitHub Actions
- Alembic

Additional Cloud

- Azure Data Factory
- Synapse
- Azure SQL
- BigQuery
- Dataflow
- Dataproc

EDUCATION

- Bachelor of Computer Applications (BCA) — Dr. Babasaheb Ambedkar Open University

CERTIFICATIONS

- **IBM Data Engineer** — IBM, 2025-01 (ID 61R9EX54ZNVG)
- **IBM Data Warehouse Engineer** — IBM, 2025-01 (ID 04EJQFZ6OU7U)
- **dbt Fundamentals** — dbt Labs, 2025-02 (ID 134685079)
- **Data Engineering Professional Certificate (V2)** — Coursera, 2025-01 (ID H1LOTACMSSLG)
- **Dagster & dbt** — Dagster Labs, 2024-11 (ID scmdubikqm)
- **UAE AI Camp Certificate 2024** — Artificial Intelligence Office (UAE), 2024-07
- **Certified Prompt Engineering** — Promptmaster, 2024-05