DATA ENGINEER CAREER TRACK – 2025 EDITION

" The architects of the data-driven future."

Market Demand Note

Data Engineers are the backbone of modern data-driven organizations, enabling analytics, machine learning, and decision-making by **designing, building, and maintaining scalable data pipelines**.**Duration**: 4–5 Months | **Mode**: Online/Offline

📘 Data Engineer Career Track — 2025 Edition

Table of Contents

Module 1: Introduction to Data Engineering (6 Hours)

Learning Objectives:
Understand the data engineer’s role, the data ecosystem, and pipeline design basics.

Topics Covered:

- Data Engineer vs. Data Scientist vs. ML Engineer: Responsibilities, skills, and career paths in the Bangalore tech ecosystem.
- Data Pipeline Architecture: End-to-end data flow from ingestion to analytics, including source systems, ingestion, storage, processing, and serving.
- Batch vs. Streaming Data: When to use each, latency requirements, real-world Bangalore use cases (e-commerce, fintech, IoT).
- Data Types: Structured (SQL tables), semi-structured (JSON, XML), unstructured (images, logs), and their processing needs.
- Data Engineering Tools: Open-source stack overview (Spark, Airflow, Kafka, etc.) and local adoption trends.

Practical Exercise:
Draw a data flow diagram for an Indian e-commerce analytics system, highlighting batch and streaming components, data sources, and analytics endpoints.

Module 2: Python for Data Engineering (16 Hours)

© 2025 Udaan Tech Academy. All Rights Reserved.

Learning Objectives:
Master Python for ETL, data wrangling, and workflow automation.

Topics Covered:

- Python Basics: Variables, functions, classes, modules, error handling, logging.
- Data Structures: Lists, dicts, sets, comprehensions, generators.
- Pandas Mastery: DataFrames, series, indexing, groupby, pivoting, merging, handling missing data.
- File I/O: Reading/writing CSV, JSON, Parquet, Avro files; compression techniques.
- Database Connectivity: SQLAlchemy, PyODBC, async database access.
- APIs & Web Scraping: Requests, beautifulsoup4, async HTTP clients.
- Performance Optimization: Vectorization, multiprocessing, memory management.

Practical Exercise:
Write a Python pipeline that extracts product data from a Bangalore-based e-commerce API, cleans/transforms it, and loads it into PostgreSQL. Add logging, error handling, and performance profiling.

Module 3: Databases & Data Modeling (14 Hours)

Learning Objectives:
Design, optimize, and administer relational and NoSQL databases for analytics.

Topics Covered:

- Relational Databases: PostgreSQL, MySQL—installation, administration, psql/mysql CLI, backup/restore.
- NoSQL Databases: MongoDB basics, collections, queries, indexing, aggregation; Cassandra for time-series data.
- Data Modeling: Star schema, snowflake schema, fact/dimension tables, slowly changing dimensions.
- Indexing & Query Optimization: Explain plans, query tuning, partitioning, materialized views.
- Database Internals: Transactions, isolation levels, MVCC, locking.

© 2025 Udaan Tech Academy. All Rights Reserved.

Practical Exercise:
Create a star schema for a Bangalore retail chain's sales analytics platform. Ingest sample data, write complex analytical queries, and optimize performance.

Module 4: Data Warehousing Concepts (10 Hours)

Learning Objectives:
Understand OLAP, partitioning, and modern cloud warehousing.

Topics Covered:

- OLTP vs. OLAP: Transactional vs. analytical workloads, schema design differences.
- Cloud Warehouses: Concepts only—BigQuery, Redshift, Snowflake architecture; compare with on-premise.
- ETL vs. ELT: When to transform before or after loading.
- Partitioning & Clustering: Range, list, hash partitioning; impact on query performance.
- Data Lake vs. Data Warehouse: Use cases, integration patterns.

Practical Exercise:
Design a warehouse schema for a streaming platform (e.g., Hotstar Asia) with user, content, and engagement data. Write ETL/ELT logic for daily batch ingestion.

Module 5: Big Data Ecosystem (12 Hours)

Learning Objectives:
Set up and query distributed data storage and processing systems.

Topics Covered:

- Hadoop Architecture: HDFS, NameNode, DataNode, YARN.
- Hive & Impala: SQL-on-Hadoop, external vs. managed tables, querying Parquet/ORC.
- Optimized Storage: Parquet, ORC file formats, schema evolution, compression.
- Data Lake Patterns: Raw, curated, and serving zones; metadata management.

Practical Exercise:
Store large web server logs into HDFS, create Hive tables over Parquet, and run analytical queries. Compare performance vs. traditional RDBMS.

Module 6: Data Processing with Spark (20 Hours)

Learning Objectives:
Process, clean, and analyze massive datasets with PySpark.

Topics Covered:

- Spark Core: RDDs, lazy evaluation, transformations & actions, caching.
- DataFrames & Spark SQL: Schema inference, UDFs, window functions, joins.
- PySpark ETL: Reading/writing various formats, handling corrupted data, incremental processing.
- Performance Tuning: Partitioning, broadcasting, skew handling.
- Integration: Connecting Spark to databases, cloud storage, Kafka.

Practical Exercise:
Build a PySpark job to clean, aggregate, and analyze web logs from an Indian SaaS company. Optimize for performance and write output to Parquet.

Module 7: Data Pipelines & Orchestration (16 Hours)

Learning Objectives:
Design, schedule, monitor, and troubleshoot complex data workflows.

Topics Covered:

- Airflow Fundamentals: DAGs, operators, sensors, XComs, task dependencies.
- Monitoring & Alerting: SLA misses, retries, task logging.
- dbt: Data transformation layer, testing, docs, version control.
- CI/CD for Data Pipelines: Automated testing, deployment strategies.
- Failure Handling: Idempotency, backfill, SLA management.

Practical Exercise:
Create an Airflow DAG to extract Bangalore weather data via API, transform, validate, and load into a data warehouse. Add dbt models for business metrics.

Module 8: Streaming Data Systems (14 Hours)

© 2025 Udaan Tech Academy. All Rights Reserved.

Learning Objectives:
Build and operate real-time data pipelines.

Topics Covered:

- Kafka Fundamentals: Brokers, topics, partitions, producers, consumers, consumer groups.
- Fault Tolerance: Replication, ISRs, delivery semantics (at least once, exactly once).
- Spark Structured Streaming: Micro-batch vs. continuous, windowing, joins, watermarking.
- Real-Time Dashboards: Superset, Power BI (Community Edition) live connections.
- Use Cases: Fraud detection, recommendation engines, IoT analytics.

Practical Exercise:
Create a Kafka pipeline to ingest live social media posts (Twitter/Reddit India), process with Spark Structured Streaming, and visualize trending topics in Superset.

Module 9: Cloud Data Engineering (16 Hours)

Learning Objectives:
Architect, build, and optimize data platforms in the cloud.

Topics Covered:

- Cloud Storage: S3, GCP Storage, Azure Blob—consistency models, cost optimization.
- Managed Big Data: Dataproc, EMR, HDInsight—provisioning, autoscaling, spot instances.
- Cloud ETL: AWS Glue, GCP Dataflow, Azure Data Factory—serverless job orchestration.
- Data Lakehouse: Delta Lake, metadata layers, schema enforcement, time travel.
- Cost Management: Monitoring, tagging, rightsizing.

Practical Exercise:
Build an end-to-end data pipeline on AWS: Ingest IoT sensor data into S3, process with EMR Spark, serve analytics via Superset, and optimize for cost/performance.

© 2025 Udaan Tech Academy. All Rights Reserved.

Module 10: Data Governance & Security (8 Hours)

Learning Objectives:
Implement data quality, access control, and compliance.

Topics Covered:

- Data Quality: Validation rules, anomaly detection, lineage tracking.
- Access Control: Row-level security, column masking, RBAC implementations.
- Privacy & Compliance: GDPR, CCPA concepts, data masking, tokenization.
- Metadata Management: Data catalog tools, discovery, lineage.

Practical Exercise:
Add data quality checks to an existing ETL pipeline. Implement row-level security in PostgreSQL. Document lineage for a critical dataset.

Module 11: Capstone Projects (20 Hours)

Choose one or combine:

- Real-Time Fraud Detection: Kafka + Spark Streaming + Dashboard (banking/fintech context).
- E-Commerce Data Lake: Ingest, clean, and serve product/customer data for analytics (Flipkart/Amazon India scenario).
- IoT Sensor Data Platform: Stream processing and storage for predictive maintenance (Bengaluru smart city/industry 4.0).
- Social Media Trend Analysis: Real-time sentiment dashboard for Indian languages.
- Healthcare Data Warehouse: HIPAA-inspired pipeline for patient analytics (with synthetic data).

Deliverables:
GitHub repo, architecture diagrams, deployed demo, technical docs, business impact summary.

Module 12: Career Preparation (8 Hours)

Topics Covered:

© 2025 Udaan Tech Academy. All Rights Reserved.

- Portfolio Building: Showcase capstone projects with READMEs, live demos, video walkthroughs.
- Resume Optimization: Highlight data pipeline, cloud, and optimization skills. Quantify impact.
- LinkedIn & Networking: Engage with Bangalore data engineering groups, attend meetups/webinars.
- Interview Prep: SQL puzzles, system design (data pipelines, scaling, trade-offs), case studies.

Practical Exercise:
Mock interviews, peer code reviews, portfolio feedback sessions.

© 2025 Udaan Tech Academy. All Rights Reserved.