Data Engineering

Duration – 100 Hours (6 Months)

Phase 1: Foundations (10 hours)

- **SQL** (most important skill)
  - Basics → Intermediate → Advanced
  - Joins, subqueries, CTEs, Window functions, Aggregations
  - Query optimization, Indexes, Execution plans
  - Practice: PostgreSQL / MySQL
- **Python for Data Engineering**
  - Core Python (functions, OOP basics, error handling)
  - File handling (CSV, JSON, Parquet)
  - Pandas / Polars for data manipulation
  - Working with APIs, Requests library
  - Writing clean, reusable scripts
- **Supporting Skills**
  - Git & GitHub (version control)
  - Linux / Command Line basics
  - VS Code / IDE setup

Project: Build a simple Python script that pulls data from a public API → cleans it → loads into PostgreSQL.

Phase 2: Databases, Data Modeling & Warehousing (12 hours)

- Relational Databases (PostgreSQL deep dive)
- Data Modeling
  - Normalization vs Denormalization
  - Star Schema & Snowflake Schema
  - Fact & Dimension tables
  - Slowly Changing Dimensions (SCD Type 1, 2)
- Data Warehouses vs Data Lakes vs Lakehouse

AA/19, DESBANDHU NAGAR, KOLKATA, North 24 Parganas, West Bengal, 700059
Phone: 8282884601
Email id: partner@cronusconsultants.co.in

- Modern warehouses: Snowflake / BigQuery / Redshift (pick one)
- Basic NoSQL concepts (MongoDB)

Project: Design a star schema for an e-commerce or sales dataset and implement it.

Phase 3: ETL/ELT Pipelines & Orchestration (12 hours)

- ETL vs ELT concepts
- **dbt** (Analytics Engineering) — highly recommended
    - Models, tests, documentation, macros
- **Apache Airflow** (or Prefect / Dagster)
    - DAGs, Operators, Scheduling, Retries, Monitoring
- Data Quality & Testing
- Docker basics (containerization)

Project: End-to-end ELT pipeline
API/CSV → Cloud Storage → Warehouse → dbt transformations → Airflow orchestration

Phase 4: Big Data & Distributed Processing (12 hours)

- Apache Spark (PySpark)
    - DataFrames, Transformations, Actions
    - Spark SQL
    - Performance tuning basics
- Introduction to Hadoop ecosystem (optional but useful)
- File formats: Parquet, ORC, Avro
- Batch processing patterns

Project: Process a large public dataset (e.g., NYC Taxi, Stack Overflow) with PySpark.

Phase 5: Cloud Platforms (12 hours)

Pick **one** major cloud (AWS preferred for most jobs, or GCP/Azure):

AWS Core Services

- S3, Glue, Redshift / Athena
- Lambda, IAM, CloudWatch
- EMR (for Spark)

GCP Alternative: BigQuery, Cloud Storage, Dataflow, Composer
Azure Alternative: Synapse, Data Factory, Databricks

- Infrastructure as Code basics (Terraform optional but valuable)
- CI/CD for data pipelines

Project: Full cloud pipeline (ingest → transform → load) running on a schedule

Phase 6: Streaming, Advanced Topics & Production Skills (12 hours)

Goal: Move beyond batch

- Real-time / Streaming
    - Apache Kafka basics
    - Spark Streaming or Flink (introduction)
- Data Observability & Monitoring
- Data Governance, Security & Access Control
- System Design for Data Platforms
- Cost optimization

Phase 7: Hands on Project (30Hours)