Study plan

1. Identify Skill Enhancement Area

- **Python**: Variables, loops, functions, exception handling, file I/O, JSON, Requests, Pandas, task automation.
- **SQL**: SELECT, WHERE, JOIN, GROUP BY, window functions, nested queries, query optimization, schema design.
- **Hadoop**: HDFS concepts, file commands, MapReduce basics, hands-on with small datasets.
- **Spark**: PySpark DataFrames, RDDs, Spark SQL, transformations, aggregations, performance tuning.
- Data Tools: Airflow DAG basics, Kafka streaming architecture, cloud storage & data warehousing (AWS S3, Redshift, BigQuery).

2. Specific Learning Objectives

- Python:
* Automate CSV cleaning and API data fetching.
* Use Pandas for data wrangling and Requests/JSON for API integration.

- SQL:
* Write optimized analytical queries.
* Build a simple dashboard using a public dataset (e.g., ecommerce).

* Optimize queries and design schemas.

- Hadoop:
* Store and retrieve data on HDFS.
* Run a simple MapReduce job and analyze logs.

- Spark:
* Build a PySpark pipeline: filtering, joining, aggregating.
* Use Spark SQL to query large datasets.

- Other Tools:
* Schedule ETL jobs with Airflow.
* Stream sample data using Kafka.

* Deploy an end-to-end pipeline on a cloud platform.

3. Select Resources and Courses

Python:

- [Python for Data Engineering – DataCamp/Coursera]
- Hackerrank Python Challenges (daily practice).

SQL:

- Mode Analytics SQL Tutorial
- LeetCode SQL Problems (easy → hard).

Hadoop:

- Simplilearn or Edureka Hadoop Beginner YouTube series.
- *Hands-on Hadoop with Big Data* (Udemy).

Spark:

- *Big Data with PySpark* (freeCodeCamp on YouTube).

Other Tools:

- Airflow: Astronomer.io Academy
- Kafka: *Apache Kafka Series* (Udemy).
- Cloud: AWS Skill Builder (Redshift, S3), Azure Data Engineering labs.

4. Schedule Study Sessions

Weekdays (Mon–Fri): ~2 hrs/day
* 1 hr: Python/SQL Practice (alternate days)
* 1 hr: Video lectures or platform practice

Weekends (Sat–Sun): 8+ hrs/day
* 3 hrs: Hadoop/Spark
* 2 hrs: Project Work or Practical Exercises
* 2 hrs: Airflow/Kafka/Cloud
* 1 hr: Review and Reflection

Total Study Time per Week: ~26 hrs
Total Over 2 Months: ~208+ hrs

5. Incorporate Practical Application

- Weekly Mini Projects:

* Week 2: Python script to clean and analyze CSV data.
* Week 4: SQL report/dashboard using public dataset.
* Week 5: Store data in HDFS and run a MapReduce job
* Week 6: PySpark job: filter, join, aggregate large dataset.
* Week 7: Simple Airflow DAG for ETL pipeline automation.

* Week 8: End-to-end pipeline: cloud storage → Spark → summary output.

- Showcase Work:

- **Weekly**: Note what worked well and gaps to revisit.

- **Monthly**: Revisit weaker topics (e.g., Spark SQL optimization) and adjust next steps.

- **End of Plan**: Create a polished GitHub portfolio, highlighting your Python, SQL, and big data skills.