Limited scheduling:

Before the Apache Airflow we need to schedule thee data pipelines using the cron jobs or custom scripts. There is lack of flexibility in defining the complex dependencies and workflow.

Mounitoring and Alerting:

Mounitoring the dealth and progress of the data pipeline was a manual process,making it difficult to identify the risks.

Tracking:

Tracking the origin and flow of data through pipeline was cahllening,hindering(making it to difficult) to troubleshooting and data quality checks.

Airflow will replace the custom python scripts with the directed Acyclic Graphs (DAGS).

This visual workflow will define the tasks and show their dependencies.

Easy maintenance: backs provide easy to define workflow and maintain, collaboration.

Scalability: DAGS easily ed to accommodate to the growing data volumes and complex workflows .

Flexible scheduling:

* Apache Airflow offers the robust scheduling
  + **Time based scheduling**:run at regular intervals (daily hourly etc).
  + **Event-driven Triggers**:tasks(pipeline) will be trigurredd by the external events or data avaliabillity.
  + **Dependencies**: airflow will ensure the tasks will run on the correct order based on the defined dependencies.

Mounitoring and Alerting :

* Apache airflow will provide the ui to mounitor your pipelies.
  + We can view:
    - Status of the task
    - Identify Errors
    - Track progress

We can iintegrate the alerting systemto notigy any issues

Tracking:

Apache Airflow

Task:

Normally we will write a script to extract the data from the API. Later we will perform the transformations. And we will load the data.

Above one is one task.

Suppose we have the 2 -3 tasks means we will schedule it suing the CRON job. But in real-time we will have the thousands of tasks we need to scheduling and monitoring and alerting and tracking and scaling. etc

When the data grows we need to create more pipelines and what if you want to rul all these in the specific order. Like first data extraction should be done second transformation should be done third loading should be done.

We need to ensure that we all the teasks are execution is done on the correct order.

Managing all this using the python script is the headace.