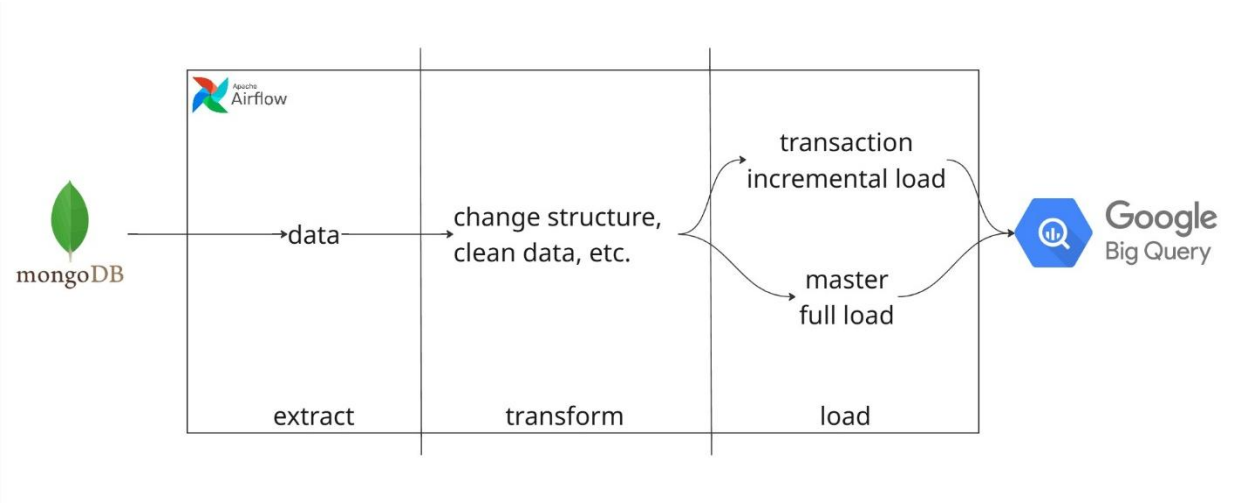
1. Pipeline



2. ออกแบบ Pipeline โดยใช้ Apache Airflow

- เราจะตั้งค่า Workflow ใน Airflow ให้ทำงานแบบ daily batch ทุกวัน
- ขั้นตอนหลักคือ
    - Extract: ดึงข้อมูลทั้ง Master Data และ Transactional Data จาก MongoDB ที่เป็นข้อมูลกึ่งโครงสร้าง (Semi-structured)
    - Transform: ทำความสะอาดข้อมูลและแปลงข้อมูลให้เป็นรูปแบบ Structured เพื่อให้ง่ายต่อการใช้งานและวิเคราะห์
    - Load: โหลดข้อมูลที่ผ่านการแปลงแล้วเข้า BigQuery โดย
        - Master Data ทำแบบ Full Load เพื่ออัปเดตข้อมูลใหม่หรือแก้ไขข้อมูลเดิม
        - Transactional Data ทำแบบ Incremental Load โดยดึงเฉพาะข้อมูลใหม่หรือที่เปลี่ยนแปลงตั้งแต่รอบก่อนหน้า เพื่อประหยัดเวลาและทรัพยากร
- ใน Airflow จะมีระบบ Monitor และ Alert แจ้งเตือนถ้าเกิดข้อผิดพลาดในแต่ละขั้นตอน เพื่อให้ทีมงานรีบแก้ไขทันที

3. ควรเก็บข้อมูลในรูปแบบ Structured เพราะผู้ใช้ส่วนใหญ่ไม่คุ้นเคยกับการจัดการข้อมูลแบบซับซ้อน และช่วยให้การเขียน SQL ง่ายและเร็วขึ้น ลดข้อผิดพลาด ทำให้การวิเคราะห์และสร้างรายงานอย่างมีประสิทธิภาพ