Big Data Analysis Pipeline

Jugal Gajjar
Saptorshee Nag
Kaustik Ranaware
Saurabh S R
Michael Womack

Introduction

The Problem We are trying to solve:

Provide an interface to track content performance across social media platforms. In this case, we used Youtube but it could be further expanded to show multiple platforms, which would be great for any company trying to manage a vast social media presence.

Goal:

- To build an end-to-end big data processing pipeline using Scala, capable of efficiently ingesting, storing, processing, and analyzing large datasets to generate actionable insights.
- Enable real-time data processing for fast decision-making, and ensure scalability and fault-tolerance for handling large-scale data.

Technology That Will Be Used

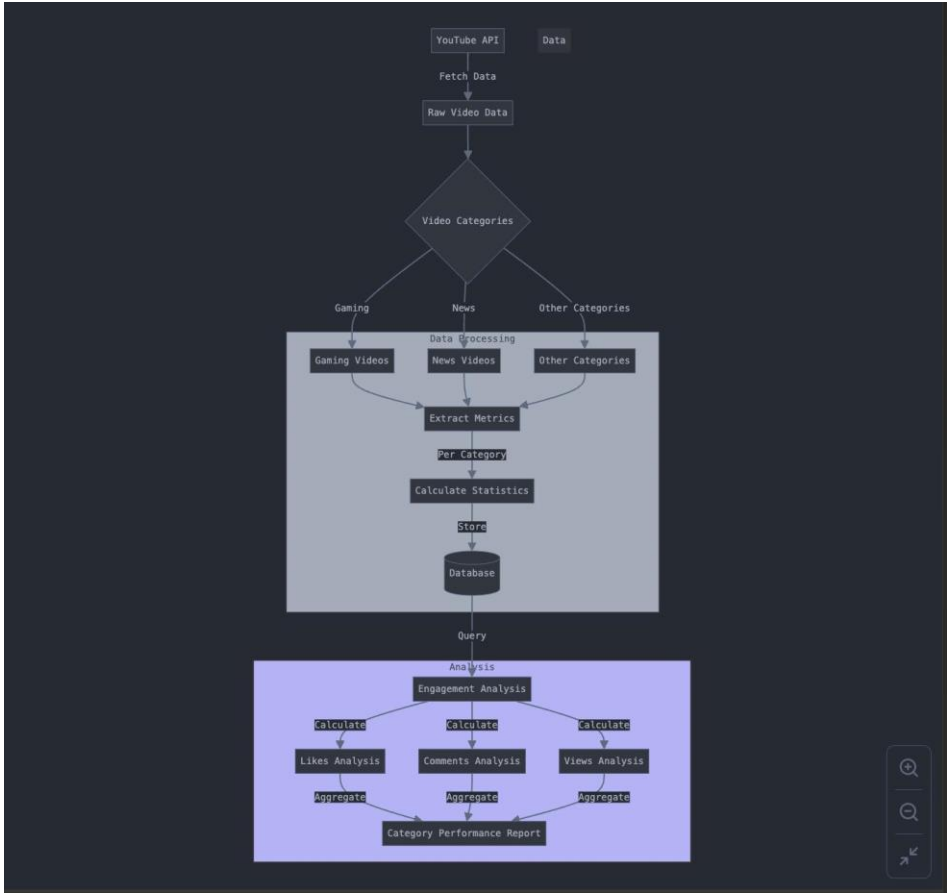
1. Scala
2. Apache Spark
3. HDFS (Hadoop Distributed File System)
4. Spark SQL/DataFrames
5. MySQL Workbench
6. Grafana

Flow Chart

Fetch Stages

Data Processing Stage

Analysis



Demo

Thank You!