ACHARYA INSTITUTE OF TECHNOLOGY

Affiliated to Visvesvaraya Technological University, Belagavi, Govt. of Karnataka.
ApprovedbyAICTE,NewDelhi

Computer Science &Engineering (DataScience)

PROJECTREPORT
ON

From BI to BigData:Explain,Design & Defend

Subject Name:Big DataAnalytics
Subject Code:BAD601

SubmittedBy : DISHA V SHETTY
1AY23CD018

Submitted To:
Ms.Surbhi

TABLE OF CONTENTS

1. About the Project

1.1 Introduction to Big Data

In today's digital world, data is constantly created through activities like video streaming, social media, online shopping, and digital payments. Each click, search, view, or interaction generates information. When this information grows very large, is produced at great speed, and comes in various formats, it is referred to as Big Data.

Big Data has five key characteristics, known as the 5 V's:

- Volume: This refers to the huge amounts of data created daily, measured in terabytes and petabytes.

- Velocity: This indicates how fast data is generated and processed in real time.

- Variety: Data comes in different forms, including structured data (like tables), semi-structured data (like logs), and unstructured data (like videos, images, and text).

- Veracity: This represents how reliable and accurate the data is.

- Value: This is about the meaningful insights derived from data that help in making decisions.

Traditional data processing systems cannot efficiently handle such large, complex, and rapidly changing data. Thus, modern Big Data technologies are necessary to store, process, and analyze this information effectively.

1.2 Overview of YouTube as a Big Data Application

YouTube is one of the world's largest big data applications because it handles massive volumes of video content, user interactions, and real-time streaming data every second

1. Huge Data Volume
Billions of videos stored and streamed.
Petabytes of data generated from uploads, views, likes, comments, and watch history.

2. Variety of Data
Video, audio, text (titles, comments), thumbnails, metadata.
User behavior data (searches, clicks, watch time).

3. High Velocity (Real-Time Data)
Continuous video uploads every minute.
Real-time recommendations and live streaming analytics.

4. Data Processing & Analytics
Uses distributed systems to store and process videos.
Machine learning analyzes viewing patterns and trends.

5. Personalized Recommendations
Suggests videos based on watch history and interests.
Improves user engagement using big data algorithms.

6. Monetization & Ads
Targets ads using user behavior data.
Measures ad performance and viewer response.

7. Content Moderation
Detects spam, harmful, or copyrighted content using AI.
Processes millions of videos automatically.

YouTube is a powerful big data application because it collects, stores, and analyzes massive, diverse, and fast-moving data to deliver personalized content, recommendations, and advertising at global scale.

1.3 Why YouTube Requires Big Data

1.Massive Video Storage & Management
YouTube stores billions of videos uploaded by users worldwide. Managing, organizing, and retrieving such huge multimedia data requires big data storage and distributed processing systems.

2.User Behavior Analysis
The platform collects vast data from user activities such as views, likes, comments, searches, and watch time. Big data analytics helps understand user preferences and viewing patterns.

3.Personalized Recommendations
Big data algorithms analyze each user's history and interests to recommend relevant videos on the homepage and suggested section, improving engagement and watch time.

4.Real-Time Streaming & Advertising
YouTube delivers millions of videos simultaneously across the world and uses big data to target ads based on user interests and measure ad performance in real time

1.4 Limitations of Traditional Business Intelligence

Traditional Business Intelligence (BI) systems are designed to analyze structured, historical data from databases and generate reports or dashboards for decision-making. However, with the growth of big data, traditional BI faces several challenges such as handling massive data volumes, processing unstructured data, and providing real-time insights. As modern organizations generate diverse and fast-moving data, BI tools alone are often insufficient for advanced analytics and predictive decision-making.

Key Limitations:

- Handles mainly structured data (tables, spreadsheets, databases).
- Limited ability to process very large-scale (big data) datasets.
- Mostly batch processing; lacks real-time analytics.
- Focuses on past trends rather than predictive or AI-driven insights.
- Static reports and dashboards can slow decision-making

2. Tools and Technologies Used:

2.1 Apache Hadoop

Apache Hadoop is an open-source framework used to store and process massive volumes of video and user data across distributed clusters. It enables scalable and fault-tolerant big data processing for YouTube.

Hadoop mainly consists of:

- HDFS (Hadoop Distributed File System) – Stores large video files, logs, and metadata by distributing them across multiple servers.
- MapReduce – Processes huge datasets such as watch history, clicks, and viewing patterns in parallel.

For YouTube, Hadoop stores user activity logs, video metadata, and streaming data at scale.

2.2 Apache Spark

Apache Spark is a fast in-memory data processing engine used for real-time analytics and large-scale data processing.

Key features:
- In-memory processing
- Real-time analytics
- Batch and streaming support
- Machine learning (MLlib)

For YouTube, Spark analyzes user behavior, detects trending videos, and powers the recommendation system.

2.3 NoSQL Databases (Bigtable / Cassandra)

NoSQL databases store large-scale, unstructured, and semi-structured data across distributed systems.

Unlike traditional databases:
- No fixed schema required
- High scalability
- High availability

YouTube uses NoSQL databases to store user profiles, comments, likes, subscriptions, and watch history.

2.4 Business Intelligence & Analytics Tools

BI tools help visualize and analyze processed data for insights and decision-making.

Examples:
* Google BigQuery
* Tableau

These tools create dashboards showing:
* Daily active users
* Most viewed videos
* Watch time and engagement metrics
* Ad performance

2.5 Machine Learning

Machine Learning is used to analyze massive user and video data to automatically improve content recommendations, search results, and platform safety on YouTube.

Key uses:
* Personalized video recommendations
* Search ranking optimization
* Spam and harmful content detection
* Ad targeting and performance prediction

For YouTube, machine learning models learn from watch history, clicks, likes, and viewing time to deliver relevant videos and maintain platform quality

3. Architecture Design

3.1 Traditional Data Warehouse Architecture

In traditional systems, data is stored in relational databases and processed using structured query methods for reporting and analysis in YouTube

.

Architecture Flow – Traditional Data Warehouse

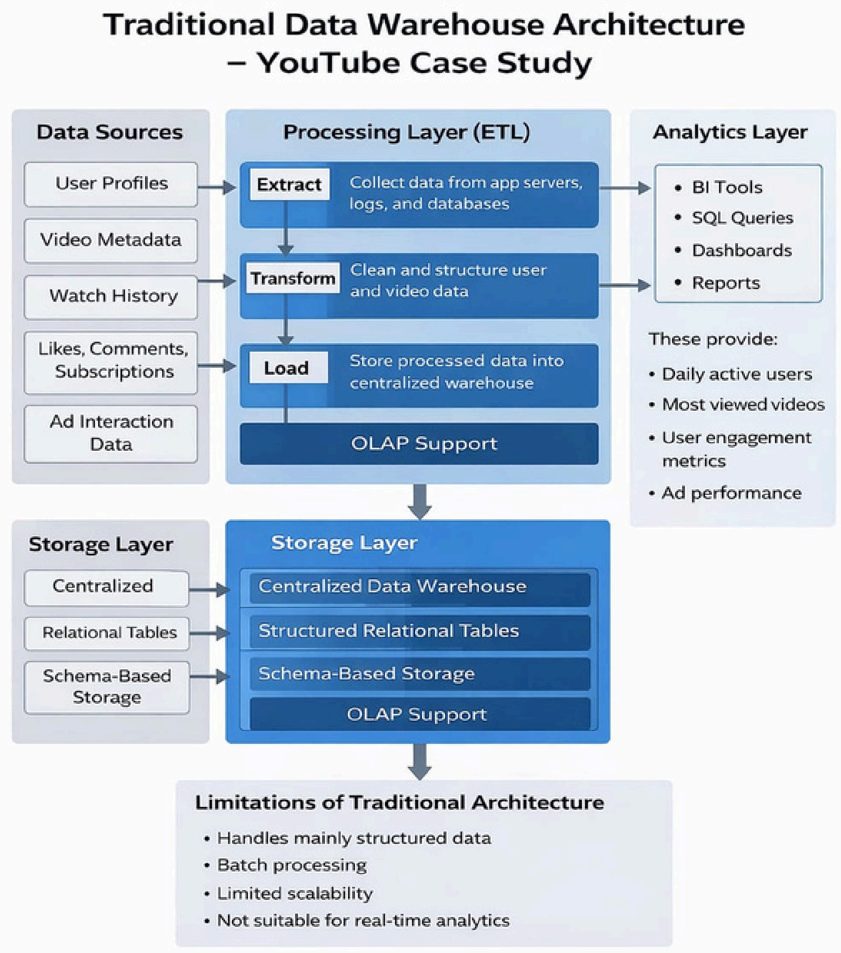


Fig 3.1 – Architecture Design

1.Data Sources

For YouTube, data sources include:

* User profiles

* Video metadata

* Watch history

* Likes, comments, subscriptions

* Ad interaction data

2.Processing Layer (ETL)

* Extract – Collect data from app servers, logs, and databases
* Transform – Clean and structure user and video data
* Load – Store processed data into centralized warehouse

3. Centralized Data Warehouse

* Structured relational tables
* Schema-based storage
* OLAP support for queries

4.Analytics Layer

* BI tools
* SQL queries
* Dashboards
* Reports

5.These provide insights such as:

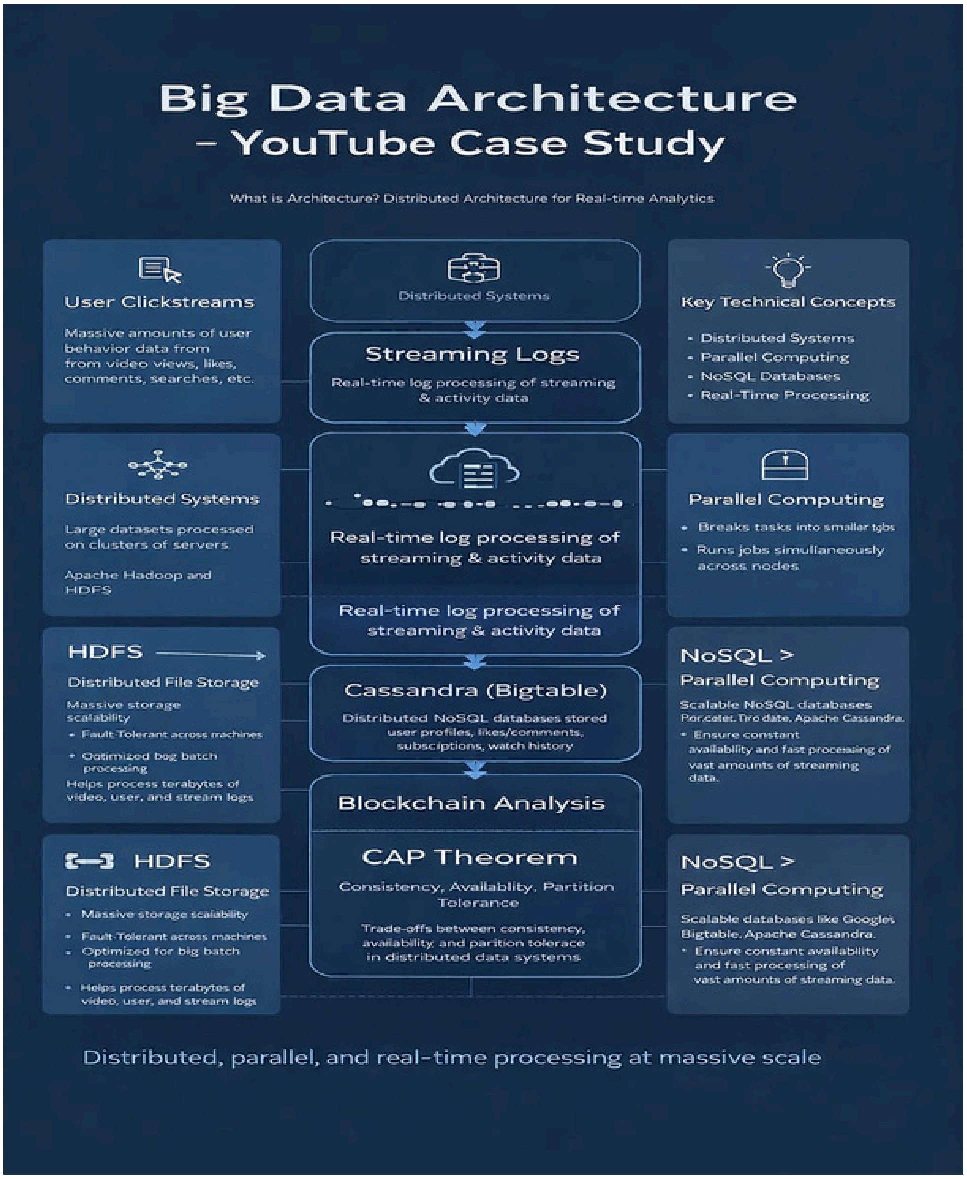
* Daily active users
* Most viewed videos
* User engagement metrics
* Ad performance

6.Limitations of Traditional Architecture

* Handles mainly structured data
* Batch processing
* Limited scalability
* Not suitable for real-time analytics

3.2 Hadoop-Based Big Data Architecture

Architecture Flow



Big Data Architecture – YouTube Case Study

What is Architecture? Distributed Architecture for Real-time Analytics

User Clickstreams

Massive amounts of user behavior data from from video views, likes, comments, searches, etc.

Distributed Systems

Streaming Logs

Real-time log processing of streaming & activity data

Key Technical Concepts

- Distributed Systems
- Parallel Computing
- NoSQL Databases
- Real-Time Processing

Distributed Systems

Large datasets processed on clusters of servers

Apache Hadoop and HDFS

Real-time log processing of streaming & activity data

Real-time log processing of streaming & activity data

Parallel Computing

- Breaks tasks into smallar tgbs
- Runs jobs simullaneously across nodes

HDFS →

Distributed File Storage

Massive storage scalability

- Fault-Tolerant across machines
- Ootimized bog batch processing

Helps process terabytes of video, user, and stream logs

Cassandra (Bigtable)

Distributed NoSQL databases stored user profiles, likes/comments, subsciptions, watch history

NoSQL >

Parallel Computing

Scalable NoSQL databases Porcatet. Tiro date. Apache Cassandra.

- Ensure constant availablity and fast processing of vast amounts of streaming data.

Blockchain Analysis

HDFS

Distributed File Storage

- Massive storage scalability
- Fault-Tolerant across machines
- Optimized for big batch processing
- Helps process terabytes of video, user, and stream logs

CAP Theorem

Consistency, Availablity, Partition Tolerance

Trade-offs between consistency, availability, and parition tolerace in distributed data systems

NoSQL >

Parallel Computing

Scalable databases like Google's Bigtable. Apache Cassandra.

- Ensure constant availablity and fast processing of vast amounts of streaming data.

Distributed, parallel, and real-time processing at massive scale

Fig 3.2 – Hadoop Architecture

1.Data Sources
For YouTube, big data sources include:
* User clickstreams (views, searches, likes, comments)
* Video uploads and metadata
* Watch history and session logs
* Streaming and buffering logs
* Ad interaction data

2.Processing Layer
* Distributed processing using Hadoop & Spark
* Real-time stream processing of user activity
* Parallel computation across clusters
* Machine learning for recommendations & trends

3.Storage Layer
* HDFS distributed file storage
* NoSQL databases (Bigtable/Cassandra)
* Data lakes for raw video & log data
* Scalable, fault-tolerant cluster storage

4.Analytics Layer
* Real-time analytics dashboards
* Recommendation engine outputs
* Trending & engagement analysis
* Ad targeting and performance metrics

This architecture enables distributed, parallel, and real-time analytics at massive scale for YouTube.

4. BI vs Big Data - Role Play

Conversation

Business Manager:
Honestly, I don't understand all this Big Data hype. We already use Excel and SQL reports. That's enough for our decisions.

Data Consultant:
Excel and SQL are excellent for structured and moderate-size data. But modern businesses generate massive and diverse data like website clicks, app usage, and social media which traditional BI tools struggle to handle.

Business Manager:
But we store all our data in databases and run dashboards. What's the limitation?

Data Consultant:
Traditional BI mainly works with structured tables sales, finance, inventory. It cannot efficiently process unstructured or semi-structured data such as images, logs, videos, or customer behavior streams.

Business Manager:
Do we really need that kind of messy data?

Data Consultant:
Yes. That "messy" data reveals real-time customer behavior what users browse, like, or abandon. It helps companies personalize marketing and predict demand.

Business Manager:

Still, our SQL queries run fine. Why change?

Data Consultant:

They run fine because your data fits on a single server. But today data grows to terabytes or petabytes. Traditional databases scale vertically (bigger machines), which becomes costly and slow.

Business Manager:

So what triggered the Big Data evolution?

Data Consultant:

Three factors: huge data volume, high speed of data generation, and many data formats. These are known as the 3Vs Volume, Velocity, and Variety.

Business Manager:
And Big Data technologies solve this?

Data Consultant:
Exactly. Technologies like Hadoop store and process massive datasets across clusters of inexpensive computers instead of one powerful server.

Business Manager:
But our relational database is reliable and structured.

Data Consultant:
Relational databases require fixed schemas. Big Data often has flexible or changing structure. That's why NoSQL databases such as MongoDB and Cassandra are used they scale easily and handle diverse data types.

Business Manager:
So are you saying BI is outdated now?

Data Consultant:
Not at all. BI is still essential for historical reporting, KPIs, and dashboards. Big Data complements BI by supplying richer and larger datasets.

Business Manager:
Can you explain the business benefit clearly?

Data Consultant:
Sure. In retail, BI shows last month's sales. Big Data analyzes live browsing and social trends to predict what customers will buy tomorrow.

Business Manager:
That would help marketing a lot.

Data Consultant:
Also operations. Logistics companies analyze GPS and sensor streams in real time to optimize delivery routes something Excel reports cannot do.

Business Manager:
So Big Data is about prediction and real-time insight?

Data Consultant:
Exactly. BI answers “What happened?” Big Data answers “What is happening now?” and “What will happen next?”

Business Manager:
I see Excel and SQL analyze the past, Big Data drives future decisions.

Data Consultant:
Right. Companies using Big Data gain faster insights, personalization, and competitive advantage.

Business Manager:
Alright, I understand now. We should explore Big Data along with our BI tools.

Data Consultant:
That’s the best approach BI for structured reporting, Big Data for scale, speed, and advanced intelligence.

5. Analytics & Tool Match

In Big Data environments like YouTube, different types of analytics are used to answer different business questions. Each analytics type uses specific tools and technologies.

Descriptive Analytics

Answers: What happened?

- Summarizes historical data using reports and dashboards.
- Identifies trendsand patterns from past data.

Diagnostic Analytics

Answers: Why did it happen?

- Analyses data to find the root cause of a problem or trend.
- Uses comparison and correlation analysis to explain changes.

Predictive Analytics

Answers: What will happen next?

- Uses historical data and machine learning to forecast future outcomes.
- Identifies patterns to predict trends and user behaviour.

Prescriptive Analytics

Answers: What action should be taken?

- Recommends the best action based on predictive results. Uses
- Optimization and AI models to improve decision-making.

Business Question	Analytics Type	Tool Used
What happened?	Descriptive Analytics	Hadoop + BigQuery +Tableau
Why did it happen?	Diagnostic Analytics	Spark + Data Mining
What will happen next?	Predictive Analytics	Spark MLlib / Machine Learning
What action should be taken?	Prescriptive Analytics	AI Recommendation Engine + NoSQL

Table 5.1 – Analytics Tool

Analytics helps convert massive YouTube data into insights. Descriptive, diagnostic, predictive, and prescriptive analytics together help understand user behavior, detect trends, predict viewing patterns, and recommend relevant content and ads.

6. Contribution & Implementation

6.1 Project Contribution

This project demonstrates how a big data architecture can be applied to YouTube to efficiently manage and analyze massive volumes of video and user interaction data. It highlights the shift from traditional BI to Hadoop-based analytics and shows how distributed storage, real-time processing, and machine learning improve recommendations, engagement analysis, and ad targeting on the platform

.

Key Contributions:

- Designed a YouTube-based big data architecture model.
- Mapped analytics types to big data tools and technologies.
- Explained limitations of traditional BI vs big data.
- Demonstrated role of ML in recommendations and trends.

6.2 Implementation Approach

The implementation follows a layered big data pipeline where YouTube data flows from sources through distributed processing and scalable storage to analytics outputs. User activity, video, and streaming data are processed using Hadoop and Spark, stored in HDFS/NoSQL systems, and analyzed using machine learning and BI dashboards for insights and decision support.

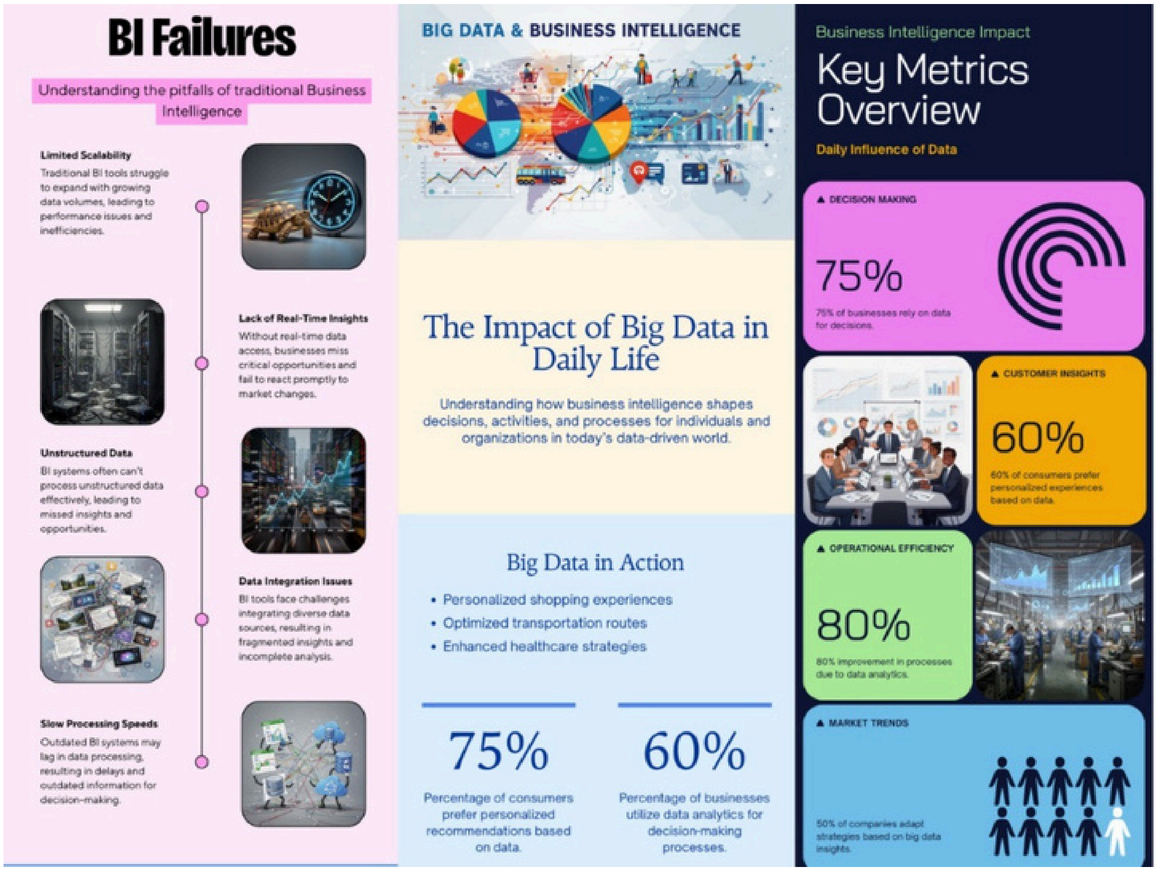
Implementation Steps:

- Collect user, video, and log data from sources.
- Process data using Hadoop and Spark.
- Store in HDFS and NoSQL databases.
- Apply ML for recommendations and prediction.
- Visualize insights using BI dashboards

7. Results

The results of this project show that Big Data architecture is more scalable and efficient than traditional BI systems for large platforms like YouTube. It effectively handles massive video data, user activity, and real-time analytics for personalized recommendations and smooth streaming

The primary outcome of this project is the comparative analysis between Traditional Data Warehouse Architecture and Hadoop-Based Big Data Architecture in the context of YouTube. The study also classified analytics into four categories: descriptive, diagnostic, predictive, and prescriptive analytics, and mapped them with suitable Big Data tools and technologies.



BI Failures

Understanding the pitfalls of traditional Business Intelligence

Limited Scalability
Traditional BI tools struggle to expand with growing data volumes, leading to performance issues and inefficiencies.

Lack of Real-Time Insights
Without real-time data access, businesses miss critical opportunities and fail to react promptly to market changes.

Unstructured Data
BI systems often can't process unstructured data effectively, leading to missed insights and opportunities.

Data Integration Issues
BI tools face challenges integrating diverse data sources, resulting in fragmented insights and incomplete analysis.

Slow Processing Speeds
Outdated BI systems may lag in data processing, resulting in delays and outdated information for decision-making.

BIG DATA & BUSINESS INTELLIGENCE

The Impact of Big Data in Daily Life

Understanding how business intelligence shapes decisions, activities, and processes for individuals and organizations in today's data-driven world.

Big Data in Action

- Personalized shopping experiences
- Optimized transportation routes
- Enhanced healthcare strategies

75%
Percentage of consumers prefer personalized recommendations based on data.

60%
Percentage of businesses utilize data analytics for decision-making processes.

Business Intelligence Impact

Key Metrics Overview

Daily Influence of Data

▲ DECISION MAKING
75%
75% of businesses rely on data for decisions.

▲ CUSTOMER INSIGHTS
60%
60% of consumers prefer personalized experiences based on data.

▲ OPERATIONAL EFFICIENCY
80%
80% improvement in processes due to data analytics.

▲ MARKET TRENDS
50% of companies adapt strategies based on big data insights.

Fig 7.1 – Infographic poster (Big data in Daily Life)

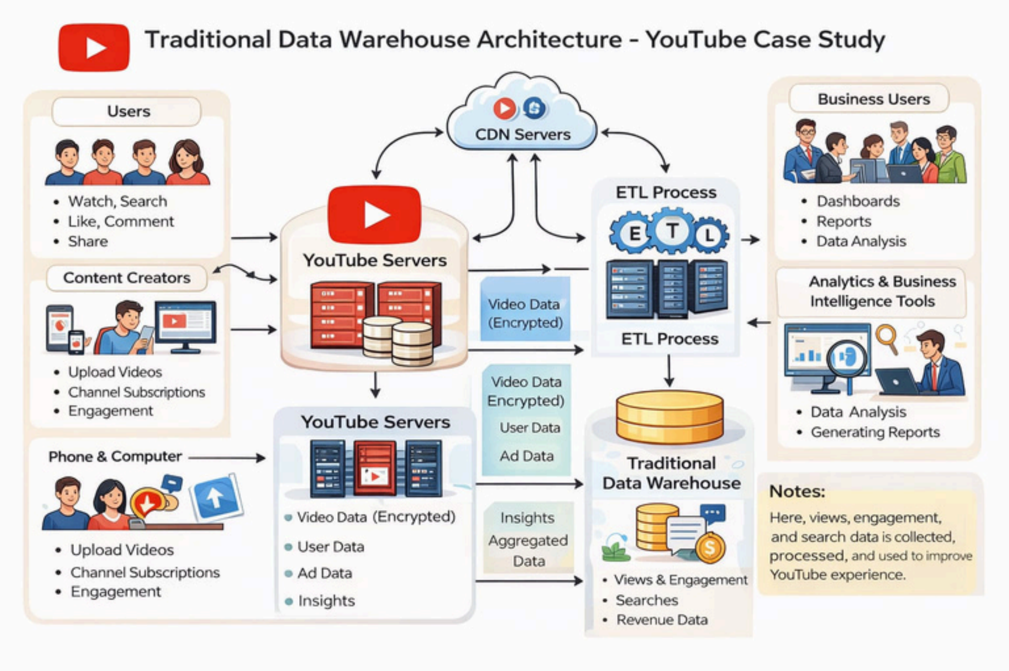


Fig 7.2 – Traditional Data Warehouse

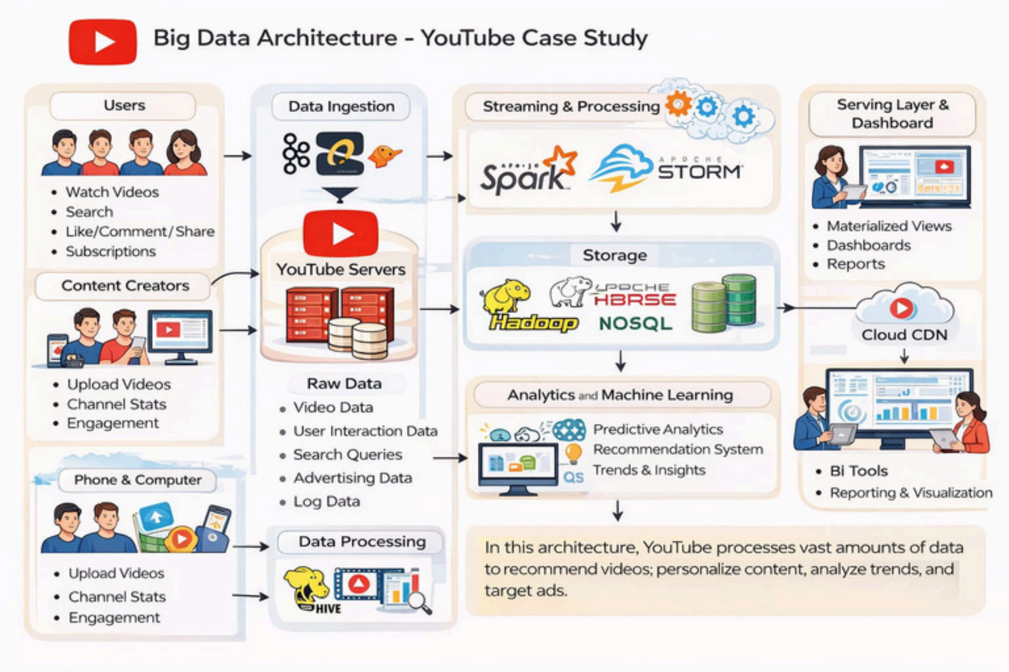


Fig 7.3 – Hadoop based Big Data

8. Conclusion

This project concludes that Big Data architecture plays a crucial role in managing and analyzing the enormous volume, velocity, and variety of data generated by modern digital platforms such as YouTube. Traditional Business Intelligence systems are limited in handling large-scale, unstructured, and real-time data, whereas Big Data technologies provide scalable storage, distributed processing, and advanced analytics capabilities required for global video streaming services.

Through this study, a comparative analysis between Traditional Data Warehouse Architecture and Hadoop-based Big Data Architecture was presented in the context of YouTube. The results demonstrated that Big Data frameworks such as distributed file systems, parallel processing engines, and machine learning models enable efficient handling of massive video uploads, user interactions, watch history, and engagement metrics. These technologies support core platform functions including video recommendation, targeted advertising, trend detection, and performance optimization.

The project also highlighted the role of different analytics types descriptive, diagnostic, predictive, and prescriptive in transforming raw platform data into meaningful insights. Real-time analytics helps YouTube personalize content feeds and improve user experience, while predictive analytics supports recommendation accuracy and audience retention. Visualization tools further assist business and content strategy decisions by presenting complex data in an understandable form.

Overall, the implementation of Big Data architecture significantly enhances scalability, speed, accuracy, and intelligence in data processing compared to traditional BI approaches. The study confirms that Big Data is not merely an improvement but a necessity for large-scale video platforms. Adopting such architecture enables platforms like YouTube to deliver seamless streaming, personalized experiences, and data-driven innovation, ensuring sustained growth and competitive advantage in the digital media ecosystem.