# ACHARYA INSTITUTE OF TECHNOLOGY

Affiliated to Visvesvaraya Technological University, Belagavi, Govt. of Karnataka.
Approved by AICTE, New Delhi

**Department of Computer Science & Engineering (Data Science)**


PROJECT REPORT
ON

# From BI to Big Data: Explain, Design & Defend

Subject Name: Big Data Analytics
Subject Code: BAD601


Submitted By:
(Sanjana H – 1AY23CD052)

Submitted To:
Ms. Surbhi

TABLE OF CONTENTS

## TASK 1: Big Data in Daily Life

The one real world Application I choose : **Instagram**

Instagram is a fantastic way to see how Big Data works in the real world. Instagram isn't just a photo app, it's a massive data factory that processes billions of interactions.

Infographic of Instagram

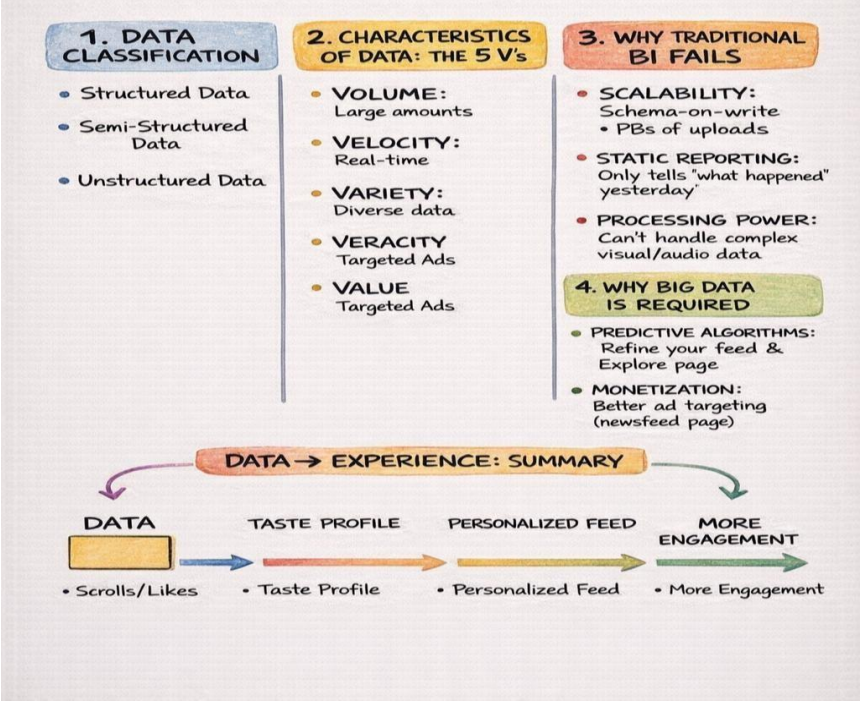


**Fig 1.1**

## Task 2 : Role play

**Manager:** Consultant, I just don't see the point. My Excel macros function, my SQL queries are clear, and my team is aware of the precise location of each data row. When something isn't broken, why fix it?

**Consultant:** It's constricted rather than broken. In a suit and tie, you're attempting to complete a marathon. You're moving, of course, but the Three V's Volume, Velocity, and variety are going to cause you to overheat.

**Manager:** My SQL server handles my volume just fine.

**Consultant:** For now. But your Traditional BI relies on Vertical Scaling buying a bigger, more expensive server every time you grow. Eventually, you hit a ceiling. Big Data uses Horizontal Scaling, adding hundreds of cheap nodes to a cluster instead.

**Manager:** Okay, but what about the Variety part? Data is data.

**Consultant:** Not in the modern world. Your SQL setup is built for Structured Data neat little tables. But 80% of your customer insights are in Unstructured Data: voice notes, images, and sensor logs. Traditional BI tools literally can't see that content.

**Manager:** So I just need a better import tool?

**Consultant:** No, you need a different philosophy. You're currently using Schema-on-Write, where you must define exactly what the data is before you save it. Big Data uses Schema-on-Read we dump everything into a Data Lake and figure out the structure only when we need to ask a question.

**Manager:** Dumping everything in a lake sounds like a mess. How do you even find anything?

**Consultant:** That's where Hadoop comes in. It uses a Distributed File System (HDFS). It breaks your data into chunks and replicates them across the cluster. It's Fault Tolerant, meaning if one computer dies, the system doesn't even skip a beat.

**Manager:** My SQL database has relationships for a reason. Everything is connected.

**Consultant:** And that's exactly why it's slow. For massive scale, we use NoSQL databases. They abandon the rigid table structure for high-speed, flexible storage that can handle millions of hits per second without locking up.

**Manager:** Is that why my real-time dashboard usually has a 24-hour delay?

**Consultant:** You're stuck in Batch Processing (ETL). By the time you see a trend, it's already historical. We need Stream Processing tools like Apache Spark to analyze data the second it's created.

**Manager:** So what's the Executive Summary? Why should I care?

**Consultant:** Because right now you're doing Descriptive Analytics . Big Data gives you Predictive Analytics. It uses Machine Learning to tell you what will happen predicting Customer Churn before the customer even knows they're unhappy.

**Manager**: Alright. You've convinced me that my spreadsheets are a safety net I've outgrown. Where do we start?

**Consultant:** We start by identifying your most valuable dark data the stuff you're collecting but not using. Once we unlock that with a modern stack, you'll stop looking at the past and start engineering the future. Let's grab a coffee and map out your first Hadoop cluster.

## TASK 3: Architecture Design Challenge

### 1.Traditional Data ware house:

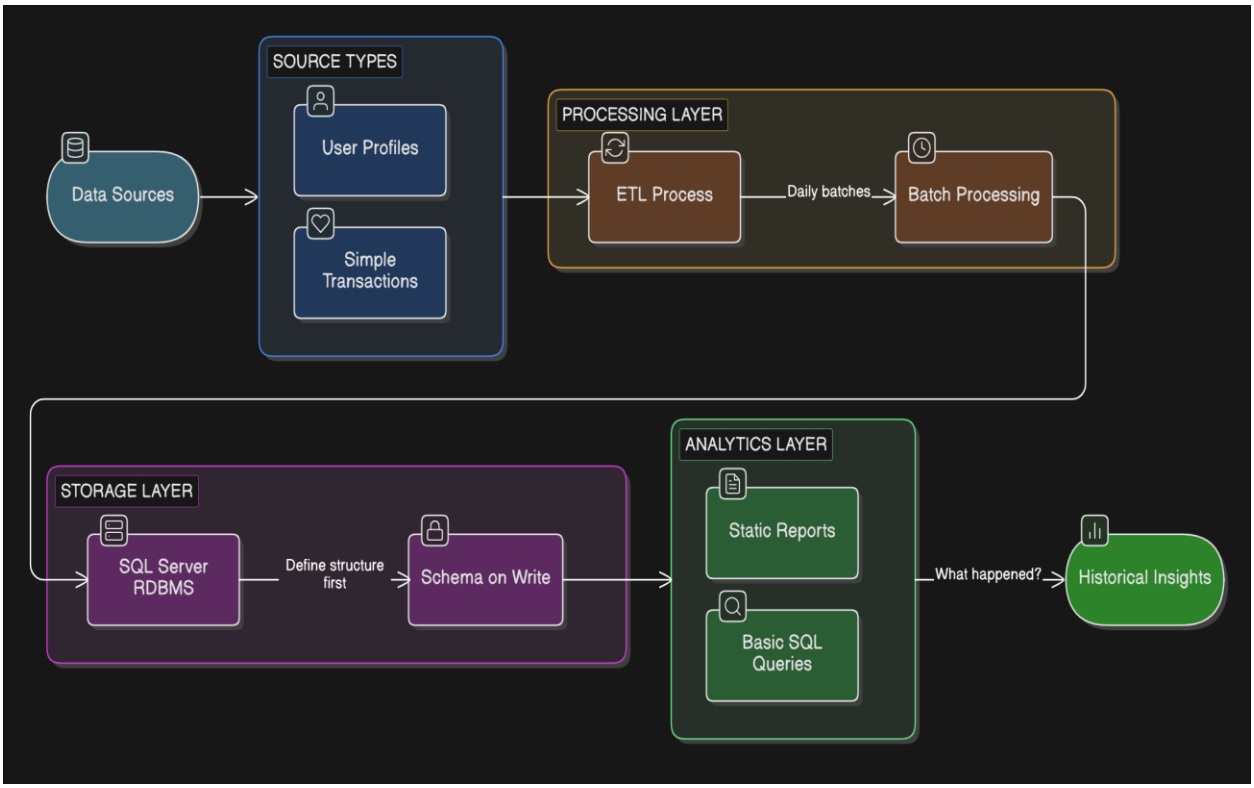


**Fig 1.2**

- **Data Sources:** Transactional databases (User login info, profile settings) and manual log entries.
- **Processing Layer: ETL (Extract, Transform, Load)**. Data is cleaned and converted into rows and columns in slow, nightly batches.

- **Storage: RDBMS** (Relational Database Management System) like Oracle or SQL Server. It uses **Schema-on-Write**, meaning the "shape" of the data is fixed.
- **Analytics Layer:** Basic **SQL Queries** and static BI dashboards

**2.Hadoop based big data architecture**

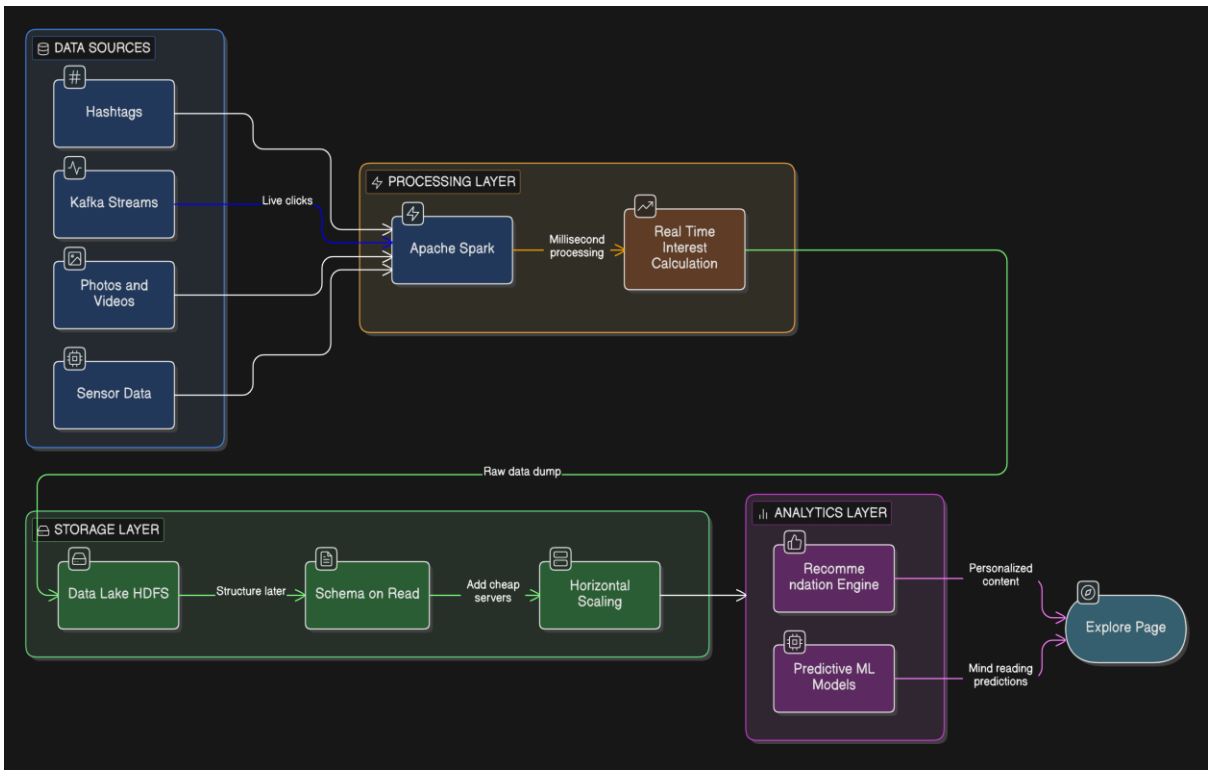


**Fig 1.3**

- **Data Sources:**

  Structured: Profile info, billing.

  Unstructured/Semi-Structured: High-res photos, Reels (Video), JSON activity logs (what you clicked), and hashtags.

- **Processing Layer:** Apache Spark: For real-time processing (updating your feed instantly).

  MapReduce: For massive overnight calculations across thousands of servers.

- **Storage**: HDFS (Hadoop Distributed File System) and NoSQL (like Cassandra). This is a Data Lake using Schema-on-Read—you store the raw data first and ask questions later.
- **Analytics Layer:** Machine Learning Models (Predicting which ads you'll click) and Recommendation Engines (The "Explore" page)

## TASK 4: Analytics & Tool Match

| Business Question | Analytics Type | Tool |
|---|---|---|
| What happened? | Descriptive Analytics | SQL, Power BI, Tableau, Excel |
| Why did it happen? | Diagnostic Analytics | Google Analytics, Drill-down Dashboards, Data Mining |
| What will happen next? | Predictive Analytics | Machine Learning (Python/R), Apache Spark, SAS |
| What action should be taken? | Prescriptive Analytics | Recommendation Engines, AI Optimization, Simulation Tools |

Table 1.1

**Bonus Challenge:**

**Explain Big Data to a 10-year-old** :

Imagine you have a toy box the size of a football stadium, and it's filled with trillions of toys some are LEGOs, some are videos, some are voice recordings, and new toys are being dumped in by a giant helicopter every single second.

That is Big Data.

It's too much for one person (or one normal computer) to ever sort through. To handle it, you need three special things:

1. The Team (Hadoop): Since the box is too big for one computer, you hire a thousand "robot helpers" to work together. Each robot takes one small corner of the stadium to organize.
2. The Speed: The robots have to work super fast because the helicopter never stops dropping new toys.

The Brain (AI): The robots are so smart they can find patterns. They might say, Hey, every time a kid picks up a blue LEGO, they usually look for a yellow wing next! In short**:** Big Data is taking a mountain of messy information and using a team of super-fast computers to find secrets or patterns that humans would never notice

Github account: SanjanaH12