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

"Healthcare Analytics — Real-Time ICU Patient Monitoring"

Subject Name: Big Data Analytics
Subject Code: BAD601

Submitted By: Chinmayi.Talawar
USN: 1AY23CD016

Submitted To:
Ms. Surbhi

Table of Contents

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)

Task 1: Big Data in Daily Life – Visual Storytelling

HEALTHCARE ANALYTICS & BIG DATA

Real-Time ICU Patient Monitoring — A Big Data Story

THE SCENARIO

A large hospital network with 500+ ICU beds generates continuous streams of patient data 24/7 — heart rate, blood pressure, oxygen levels, ventilator readings, lab results, imaging scans, and nursing notes — every second.

The question: Can traditional BI systems save lives in real time? Or do we need Big Data?

SECTION 1: DATA CLASSIFICATION IN ICU MONITORING

STRUCTURED DATA	{ } SEMI-STRUCTURED DATA	UNSTRUCTURED DATA
Neatly organized in rows & columns	**Has format but not tabular**	**No predefined structure**
• Patient ID, Age, Blood Type • Heart Rate (bpm): 72, 85, 91 • Blood Pressure: 120/80 • SpO2 %: 98, 97, 95 • Medication dosage records • Billing & insurance data	• HL7 / FHIR health records (XML) • JSON sensor logs from monitors • Ventilator event logs • Wearable device metadata • Lab result reports (PDF/JSON) • EHR system exports	• MRI / CT scan images (DICOM) • ECG waveform signals • Doctor's voice notes (audio) • Nursing narrative text notes • Radiology report PDFs • Endoscopy/surgery videos
~40% of ICU data	**~25% of ICU data**	**~35% of ICU data**

SECTION 2: THE 5 V's OF BIG DATA — Applied to ICU Monitoring

VOLUME	VELOCITY	VARIETY	VERACITY	VALUE
How MUCH data is generated	**How FAST data arrives**	**How MANY types of data**	**How ACCURATE data is**	**How USEFUL it is**
500+ ICU beds × 20+ sensors each = **10,000+ data points/sec** ~5 TB of data per day per hospital	**Every 500ms** heart monitor streams new reading **Alerts must fire in < 1 second** Delayed alarm = patient fatality risk	**1,000+ data types** Vitals, scans, text, audio, video, genomics **From 50+ different device brands** Each with different formats & protocols	**Sensor failures,** corrupted readings, duplicate entries must be filtered **~15% of raw sensor data** is noisy or erroneous	**Early sepsis detection** saves ~8,000 lives/year **Predictive alerts reduce ICU mortality by 20%** (Johns Hopkins Research, 2023)

SECTION 3 & 4: WHY TRADITIONAL BI FAILS ✕ vs WHY BIG DATA IS REQUIRED ☑

✕ WHY TRADITIONAL BI FAILS

⏱ Speed Problem

Traditional databases process batch updates (hourly or nightly). ICU monitors need millisecond responses. A patient going into septic shock cannot wait for the nightly ETL job.

Structure Problem

SQL & Excel handle only structured tabular data. They cannot store or analyze MRI scan images (DICOM), ECG waveforms, doctor voice recordings, or unstructured nursing notes.

Volume Problem

A single hospital generates 5+ TB of data per day. Excel crashes beyond ~1 million rows. Traditional data warehouses are cost-prohibitive to scale for petabytes of medical data.

Intelligence Problem

Traditional BI is backward-looking: it only tells you WHAT happened in the past via dashboards and reports. It cannot predict which patient is at risk of cardiac arrest in the next 2 hours.

☑ WHY BIG DATA IS REQUIRED

Real-Time Stream Processing

Apache Kafka ingests 10,000+ events/second from all ICU sensors. Apache Spark Streaming processes and raises alerts in under 1 second — catching deterioration before it becomes critical.

Distributed Storage (HDFS + NoSQL)

Hadoop HDFS stores all data types — structured, semi-structured, and unstructured — across commodity servers. MongoDB (NoSQL) handles JSON health records. Cassandra manages time-series sensor data at scale.

Scalable Cloud Infrastructure

AWS / Azure cloud Big Data platforms scale horizontally — add more nodes as data grows. Cost drops from $10M+ (traditional DW) to affordable pay-as-you-go cloud storage for a hospital network.

AI/ML Predictive Analytics

Spark MLlib trains machine learning models on historical patient data. These models predict sepsis risk 6 hours early, flag deterioration patterns, and recommend interventions — proactively saving lives.

REAL-WORLD IMPACT OF BIG DATA IN HEALTHCARE

20%	6 hrs	30%	$100B+
Reduction in ICU mortality with predictive analytics	Early warning time for sepsis using ML models	Faster diagnosis when using Big Data imaging analysis	Annual savings potential for US healthcare via Big Data

TASK 2 — BI vs BIG DATA: ROLE PLAY DIALOGUE
A Conversation Between a Data Consultant and a Business Manager
Application: Healthcare Analytics | ICU Patient Monitoring | 24 Exchanges

Characters:
Priya — Data Consultant. Expert in Big Data architecture and healthcare analytics.
Raj — Business Manager. 15 years of hospital operations experience. Believes Excel and SQL are sufficient.

TRADITIONAL BI LIMITATIONS

[1] Raj (Business Manager) — Opening Stance
"Priya, I honestly don't see why we need to spend so much on this Big Data project. We have Excel, SQL Server, and Power BI — that's already a solid setup. Our reports look fine. What exactly is the problem here?"

[2] Priya (Data Consultant) — The Hidden Cracks
"I understand where you're coming from, Raj. But let me ask you something simple — how long does your SQL report take when you query last year's full patient database? And what happens when you open that Excel file with two million rows?"

[3] Raj (Business Manager) — Admitting Limitations
"Okay, honestly — the SQL report sometimes takes 45 minutes and occasionally times out. And yes, Excel crashes above a million rows. But we split the files and manage it. It works, more or less."

[4] Priya (Data Consultant) — Why 'Managing' Isn't Enough
"That's the problem, Raj — 'managing somehow' is not a strategy when patient lives are involved. Our ICU generates over 5 terabytes of data every single day. Traditional BI was simply never designed for this kind of scale. Splitting Excel files is a workaround, not a solution."

[5] Raj (Business Manager) — Bigger Server Argument
"But surely we can just buy a bigger server? Upgrade the RAM, get a faster CPU? That seems simpler and cheaper than rebuilding everything from scratch."

EVOLUTION OF BIG DATA

[6] Priya (Data Consultant) — Why Bigger Hardware Isn't the Answer
"That used to be the answer — ten years ago. But the nature of data has changed completely. We're now collecting real-time ECG waveforms, MRI and CT scan images, doctor voice recordings, wearable device streams — data types that didn't even exist in our systems a decade ago. One MRI scan is 50 to 100 megabytes. Multiply that by 500 patients and multiple scans per day. No single server can keep up with this anymore, no matter how expensive it is."

[7] Raj (Business Manager) — Defending SQL
"Alright, I see the volume issue. But why not just use SQL? We've used it for 15 years, our IT team knows it inside out, and it has always been reliable. Why replace something that works?"

[8] Priya (Data Consultant) — SQL's Blind Spot

"Nobody is replacing SQL entirely — it still works for structured data. But here is its critical limitation: SQL can only handle data that fits neatly into rows and columns. It cannot store an MRI image, a doctor's voice note, or an ECG waveform. About 35 percent of everything our hospital generates is unstructured data — and SQL is completely blind to all of it."

[9] Raj (Business Manager) — The Speed Problem

"Okay, I hadn't thought about the unstructured data angle. But what about speed? Can't we just optimize our SQL queries — add indexes, hire a better database administrator?"

[10] Priya (Data Consultant) — Real-Time vs Batch Processing

"Optimization helps, but only up to a point. The deeper problem is that SQL and traditional BI work in batches — you collect data, process it later, and generate a report. But the ICU operates in real time. A patient going into septic shock right now cannot wait for tonight's batch report. Our system needs to detect danger and fire an alert in under one second. Traditional systems are architecturally incapable of that. This is exactly why Big Data evolved — the world became real-time, and the old tools didn't."

HADOOP & NoSQL EXPLAINED

[11] Raj (Business Manager) — What Is Hadoop?

"Alright, you've made a convincing case against the old approach. Now explain to me — what exactly is Hadoop? Every time I hear it in meetings I just nod along, but I genuinely don't know what it does."

[12] Priya (Data Consultant) — Hadoop Explained Simply

"Think of it this way. Traditional systems are like one very strong warehouse worker trying to move 10,000 boxes alone. Hadoop is like hiring 1,000 normal workers and splitting the boxes among them. Each worker handles a small portion simultaneously, and together they finish in minutes what one person would take days to do. Hadoop distributes both storage and processing across hundreds of ordinary, affordable computers. HDFS — the Hadoop Distributed File System — is the storage layer, and it can hold any type of data: images, audio, text, structured tables, everything in one place."

[13] Raj (Business Manager) — Why Not Just SQL?

"Okay, that makes sense for storage. But NoSQL — why do we need an entirely new type of database? What is actually wrong with the SQL databases we already have running?"

[14] Priya (Data Consultant) — NoSQL Explained

"SQL databases are like filing cabinets with fixed drawer sizes — every document must be the exact same format to fit. NoSQL is like a flexible storage unit where you can put in papers, photos, audio recordings, and oddly shaped objects — it adapts to whatever you give it. MongoDB, for example, stores flexible JSON documents — perfect for patient health records where every patient has slightly different fields. Cassandra handles time-series data brilliantly — ideal for continuous ICU sensor readings arriving every 500 milliseconds. SQL would struggle badly with both of these."

[15] Raj (Business Manager) — What About Apache Spark?

"And what is Apache Spark? I keep hearing that word alongside Hadoop. Are they the same thing or different?"

[16] Priya (Data Consultant) — Spark and Kafka Explained

"They work together but do different things. Hadoop is excellent for storing and batch-processing large data, but it is not fast enough for real-time analysis. Apache Spark is the processing engine that runs on top of Hadoop's stored data — and it is about 100 times faster because it processes data in memory rather than reading and writing to disk repeatedly. Spark is what lets us run machine learning models on live patient data and raise an alert within one second. Kafka is another tool — think of it as a super-highway that takes data streams from all 500 patients' sensors simultaneously and routes everything to storage instantly. Together the flow is: Kafka ingests, HDFS stores, Spark processes, ML models predict."

REAL BUSINESS BENEFITS

[17] Raj (Business Manager) — Show Me the Business Case

"Alright Priya, I'm beginning to understand the technical side. But I'm a business manager — I need to justify this to the board. What is the actual return on investment? What do we gain in concrete, measurable terms?"

[18] Priya (Data Consultant) — The ROI in Lives and Money

"Here are concrete numbers. Johns Hopkins Hospital implemented Big Data analytics in their ICU and achieved a 20 percent reduction in patient mortality. If our hospital has 500 ICU patients per month and we reduce mortality by 20 percent, that is 100 lives saved every month. Beyond the ethical dimension — fewer deaths means fewer legal cases, stronger hospital reputation, higher patient trust, and better eligibility for government funding and accreditation. Every single one of those translates directly into business outcomes."

[19] Raj (Business Manager) — How Does It Actually Work?

"That mortality figure is genuinely surprising. But how does the technology actually achieve that reduction? What does it do differently from what our doctors are already doing?"

[20] Priya (Data Consultant) — Predictive Analytics in Action

"Machine learning models are trained on millions of historical patient records and learn the subtle early warning patterns of conditions like sepsis, cardiac arrest, and respiratory failure — patterns too quiet and complex for human doctors to catch manually across 500 patients simultaneously. The Big Data system monitors every patient 24 hours a day and can predict sepsis up to six hours before symptoms become clinically obvious. That six-hour window is often the difference between a routine intervention and a life-threatening crisis. The doctor receives an alert, acts early, and the patient survives. Traditional BI only tells you what already happened. Big Data tells you what is about to happen — and gives you time to prevent it."

[21] Raj (Business Manager) — Cost Concern

"Everything you've said makes sense. My final concern is cost. Setting up Hadoop clusters, Spark, Kafka, NoSQL databases — surely that will cost far more than simply upgrading our existing SQL infrastructure?"

[22] Priya (Data Consultant) — Cloud Economics

"This is where the economics have completely changed. A traditional data warehouse capable of handling petabytes would have cost crores to build a decade ago. Today, cloud platforms like AWS, Azure, and Google Cloud provide the exact same Big Data capabilities on a pay-as-you-go model. We pay only for what we actually use, and the system scales automatically as our data grows. The upfront cost barrier that existed ten years ago is simply not a barrier today. Cloud-based Big Data is often cheaper in the long run than maintaining aging SQL infrastructure."

THE CONVERSION(CONCLUSION)

[23] Raj (Business Manager) — The Conversion

"Priya, I came into this conversation fully prepared to defend Excel and SQL. But you have addressed every concern I had — scale, speed, data types, cost, and patient outcomes. I think I need to support this initiative. Where do we even begin?"

[24] Priya (Data Consultant) — The Roadmap

"This is exactly the conversation every organization needs to have, Raj. We start with a data audit — understand what we are generating, where it lives, and what valuable information is currently being lost or ignored. Then we run a small pilot: deploy a Kafka, HDFS, and Spark stack in just the ICU, run it alongside our existing systems for three months, and demonstrate the results side by side. Once the board sees a real six-hour sepsis prediction working live on actual patients, the rest of the organization will follow on its own. Excel stays for small everyday reports. But for monitoring 500 lives in real time — Big Data is no longer optional."

Key Concepts Covered

1. Traditional BI Limitations — Excel row limits, SQL's inability to handle unstructured data, batch processing delays, and expensive scaling.
2. Evolution of Big Data — Explosion of data types from IoT, imaging, and wearables; shift from batch to real-time processing requirements.
3. Hadoop and NoSQL — Distributed storage with HDFS, flexible NoSQL databases (MongoDB, Cassandra), real-time processing with Apache Spark, and ingestion with Kafka.
4. Real Business Benefits — 20% ICU mortality reduction, 6-hour early sepsis prediction, affordable cloud infrastructure, and improved patient outcomes.

TASK 3 — ARCHITECTURE DESIGN CHALLENGE
Application: Healthcare Analytics — ICU Patient Monitoring

ARCHITECTURE 1 — TRADITIONAL DATA WAREHOUSE

DATA SOURCES

Hospital Info System Patient records, admissions	**Lab Systems** Blood tests, reports	**Pharmacy System** Medication records

▼

ETL PROCESS (Extract → Transform → Load)

ETL Tool (SQL Server Integration Services)
Runs as a nightly batch job — collects, cleans, and loads data into the warehouse

▼

STORAGE

Central Data Warehouse (Oracle / SQL Server)
Structured tables only — fixed schema — single central server

Operational Database Day-to-day transactions	**Data Marts** Department-specific subsets

▼

PROCESSING

OLAP Engine (Online Analytical Processing)
Pre-aggregates data into cubes for faster querying and reporting

▼

ANALYTICS LAYER

Power BI / Tableau Dashboards & charts	**Excel Reports** Manual analysis	**Management Reports** Weekly summaries

ARCHITECTURE 2 — HADOOP-BASED BIG DATA ARCHITECTURE

DATA SOURCES (All Types)

ICU Sensors & Monitors Vitals every 500ms	**Medical Imaging** MRI, CT scans (DICOM)	**EHR & Lab Systems** HL7/FHIR, JSON logs
Doctor Voice Notes Unstructured audio	**Wearable Devices** Continuous telemetry	

▼

INGESTION LAYER

Apache Kafka (Real-Time Message Streaming)
Collects 10,000+ events per second from all sensors — routes data streams to storage instantly

▼

STORAGE LAYER

HDFS — Hadoop Distributed File System
Stores ALL data types across hundreds of servers — petabyte-scale — structured, semi-structured, unstructured

MongoDB (NoSQL) Flexible JSON records	**Cassandra (NoSQL)** Time-series sensor data	**Data Lake** Raw data archive

▼

PROCESSING LAYER

Apache Spark (Real-Time) Fires alerts in under 1 second	**MapReduce (Batch)** Historical data analysis

▼

MACHINE LEARNING LAYER

Spark MLlib — Machine Learning Models
Predicts sepsis 6 hours early, detects cardiac risk — trained on millions of historical records

▼

ANALYTICS LAYER

Real-Time Dashboard Live status — all 500 beds	**Automated Alerts** Instant doctor notifications	**Hive / Tableau** Advanced reporting

TASK 4 — ANALYTICS & TOOL MATCH
Application: Healthcare Analytics — ICU Patient Monitoring

Business Question	Analytics Type	Definition	Tool / Technology	ICU Example
What happened?	**Descriptive Analytics**	Looks at past data to summarize what occurred. Shows history.	SQL, Power BI, Tableau, Hive	Patient's heart rate was 95 bpm at 2am. ICU had 12 critical cases last week.
Why did it happen?	**Diagnostic Analytics**	Digs deeper into past data to find the cause of an event.	Apache Spark SQL, Hadoop, Drill	Heart rate spiked because oxygen levels dropped due to ventilator misalignment.
What will happen next?	**Predictive Analytics**	Uses patterns in past data to forecast future events using ML.	Spark MLlib, Python (scikit-learn), TensorFlow	Patient has 78% chance of sepsis in next 6 hours based on current vitals trend.
What action should be taken?	**Prescriptive Analytics**	Recommends the best action to take based on predictions.	AI/ML Decision Engines, NoSQL, Recommendation Systems	Administer 500ml saline and alert on-call doctor immediately for Room 7 patient.

Brief Explanation of Each Analytics Type:

1. Descriptive Analytics
Answers 'What happened?' — It is the simplest form. Just summarizes past data using reports and dashboards. Example tool: Power BI showing last week's ICU occupancy.

2. Diagnostic Analytics
Answers 'Why did it happen?' — Goes one step deeper to find the root cause. Uses drill-down queries and data exploration. Example tool: Spark SQL finding why mortality spiked on Tuesday nights.

3. Predictive Analytics
Answers 'What will happen?' — Uses machine learning models trained on historical data to predict future events. Example tool: Spark MLlib predicting sepsis 6 hours before symptoms appear.

4. Prescriptive Analytics
Answers 'What should we do?' — The most advanced type. Not only predicts what will happen but also recommends the best action to take. Example: AI system telling doctors to administer specific medication immediately.

BONUS — EXPLAIN BIG DATA TO A 10-YEAR-OLD

Imagine your school library has 10 books.
Your teacher can easily find any book. That is how old computers work — small, neat, organised data. Easy.
Now imagine this —
Every person on Earth donates 1,000 books every single day. Not just books — also photos, videos, voice recordings, and random scribbles. In every language. Every second. Non-stop.
Your poor teacher gets completely overwhelmed. She cannot find anything. The library collapses.
This is exactly what happens to Excel and SQL when the world's data explodes. They simply cannot keep up.
So what do we do?
We hire one million robot librarians. Each robot handles a small section. They all work together at the same time — super fast. They can read books, watch videos, listen to songs, and look at photos — all at once. Ask them anything and they answer in less than one second.
THAT is Big Data.
Real example — YouTube:
Every minute, 500 hours of new videos are uploaded. YouTube uses Big Data to suggest the perfect next video for 2 billion people simultaneously. No single computer could ever do that.
Hospital example:
A hospital has 500 ICU patients. Each has 20 machines tracking heartbeat, oxygen, and blood pressure every half second. Big Data watches ALL 500 patients at once and warns the doctor 6 hours before a patient gets seriously sick. It is like a super smart friend who never sleeps and never misses anything.
The simple summary:
Old computers = one librarian with 10 books.
Big Data = one million robot librarians with billions of books, photos, videos, and songs.
Big Data is not just a lot of data. It is a smarter, faster, bigger way of handling information that our new world cannot live without.