CC IA-2

A Cloud-Native Platform for Privacy-Compliant Clinical Data Analytics

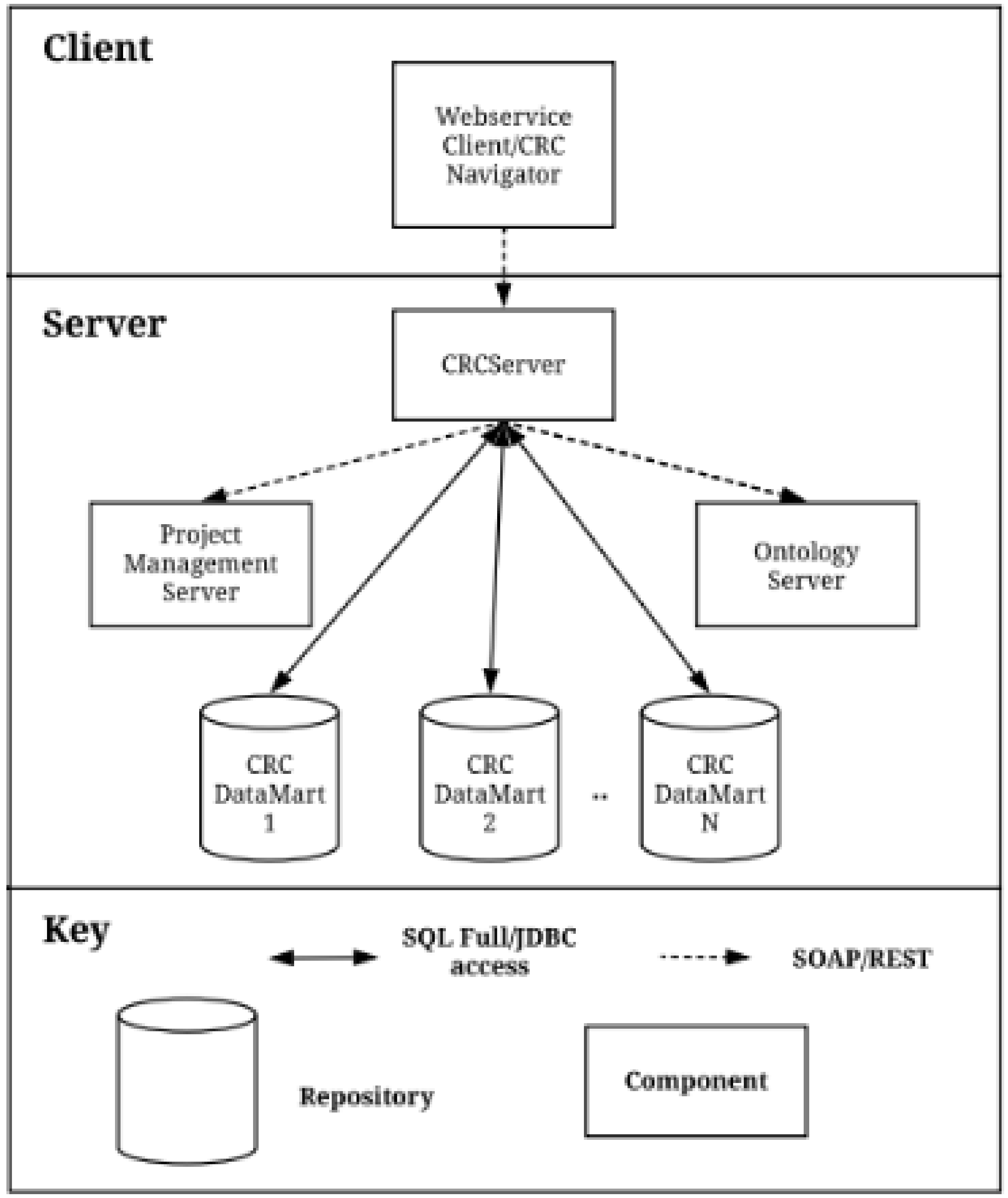
Presented by:
Smit Patil 16010139
Toyash Patil 16010122140
Khushi Poojary 16010122147

Institution: KJSSE

Date: 12|04|25

Introduction

- **Health data** is growing rapidly due to advancements in **digital health records and monitoring.**
- There's a significant need for **efficient**, **scalable**, and **privacy-compliant** analysis of this data.
- Current systems often fall short in supporting **reproducible research** across **institutions**.
- Aim of the paper: Design a **collaborative**, **open-source platform** using **cloud-native technologies for secure** and **scalable analytics.**



Abstract

- The paper presents an **architecture** combining **Apache Hadoop, Kubernetes,** and **JupyterHub.**
- Built to **support scalable, reproducible** and **flexible analysis** of large **healthcare datasets**.
- Tested with datasets exceeding **69 million patient records**.
- Prioritizes **HIPAA compliance** and **end-user usability**
- Demonstrates how **cloud-native design principles** enhance **performance** and **data security**.

Literature Survey

- **Several tools have emerged for clinical data research:**

1. **PIC-SURE**: Emphasizes querying across multiple data sources but lacks flexibility.
2. **i2b2 & SHRINE**: Focus on patient cohort exploration but do not scale well.
3. **OHDSI**: Standardizes clinical vocabularies, limited in handling unstructured data.
4. **UlTraMan & WaveformECG**: Target unstructured waveform analysis but lack scalability.

- **Most systems struggle with:Integrating structured and unstructured data.**

1. Supporting reproducible and flexible analysis.
2. Scaling to very large datasets.

Motivation

- The increasing **volume** and **complexity** of **healthcare data** — such as **EHRs, imaging** and **waveform data** — demands advanced infrastructure for meaningful **analysis**

- Current tools are fragmented and don't scale well for **real-time, privacy-sensitive, multimodal clinical research.**

- A **reproducible, secure** and **collaborative analytics framework** is necessary to accelerate **clinical insights** and support **data-driven decisions.**

Scope of the Study

Develop a cloud-native clinical analytics platform that can:

- Handle large-scale, heterogeneous datasets.
- Ensure privacy compliance and data governance.
- Support parallel processing and real-time queries.
- Provide collaborative environments via JupyterHub.
- Integrate open-source components like Apache Spark, Ceph, and Kubernetes for cost-effective deployment.

Methodology

1) Infrastructure Layer:

- Cloud VM clusters with large memory and fast storage.
- Uses Kubernetes for orchestrating services.

2) Storage Layer:

- Apache Hadoop's HDFS for structured data.
- Ceph object store for waveform data (e.g., ECG).

3) Compute Layer:

- Spark for distributed query execution.
- HIPI and Hadoop for waveform processing.

4) Service Layer:

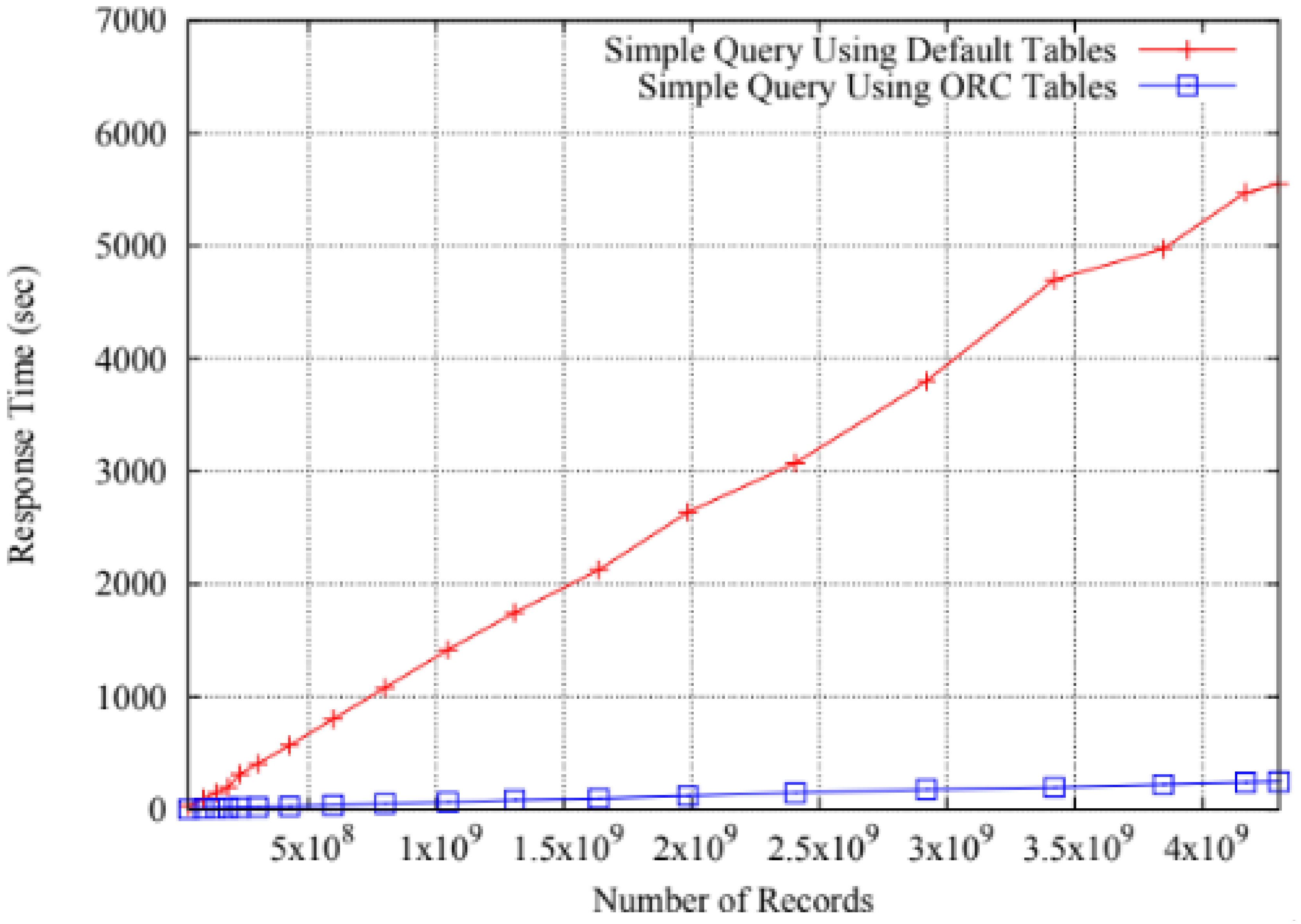
- JupyterHub for user interaction.
- Docker containers ensure reproducibility.
- CILogon & Rook manage identity and storage integration.

Observations

1) **Evaluated on two main types of queries:**
   - Simple Query: Counting lab test types across a dataset.
   - Complex Query: Joining lab results with diagnosis data to study disease patterns.

2) **Spark executor count affected performance.**

3) **Response time was improved when:**
   - Using ORC file format over default Hive tables.
   - Using higher numbers of Spark executors (to a limit).

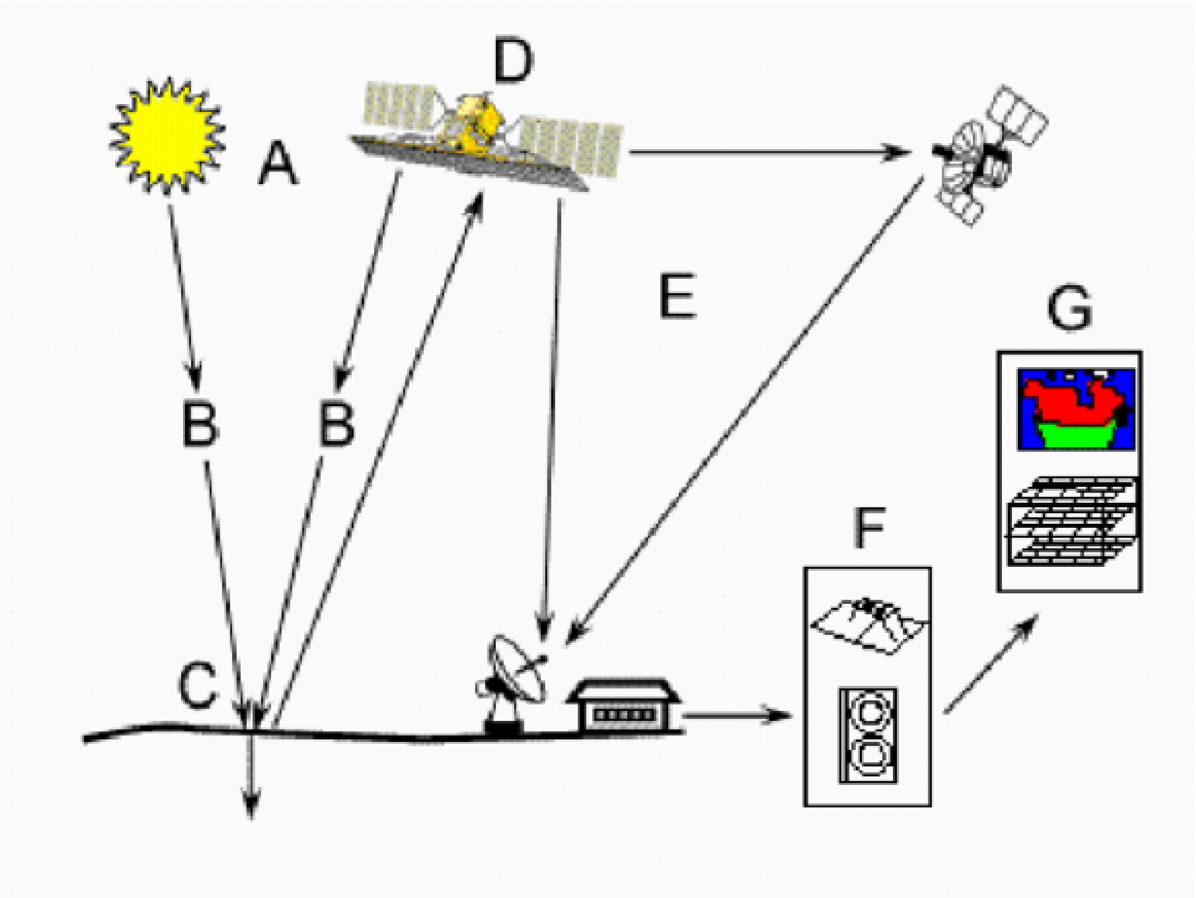


Results & Conclusion

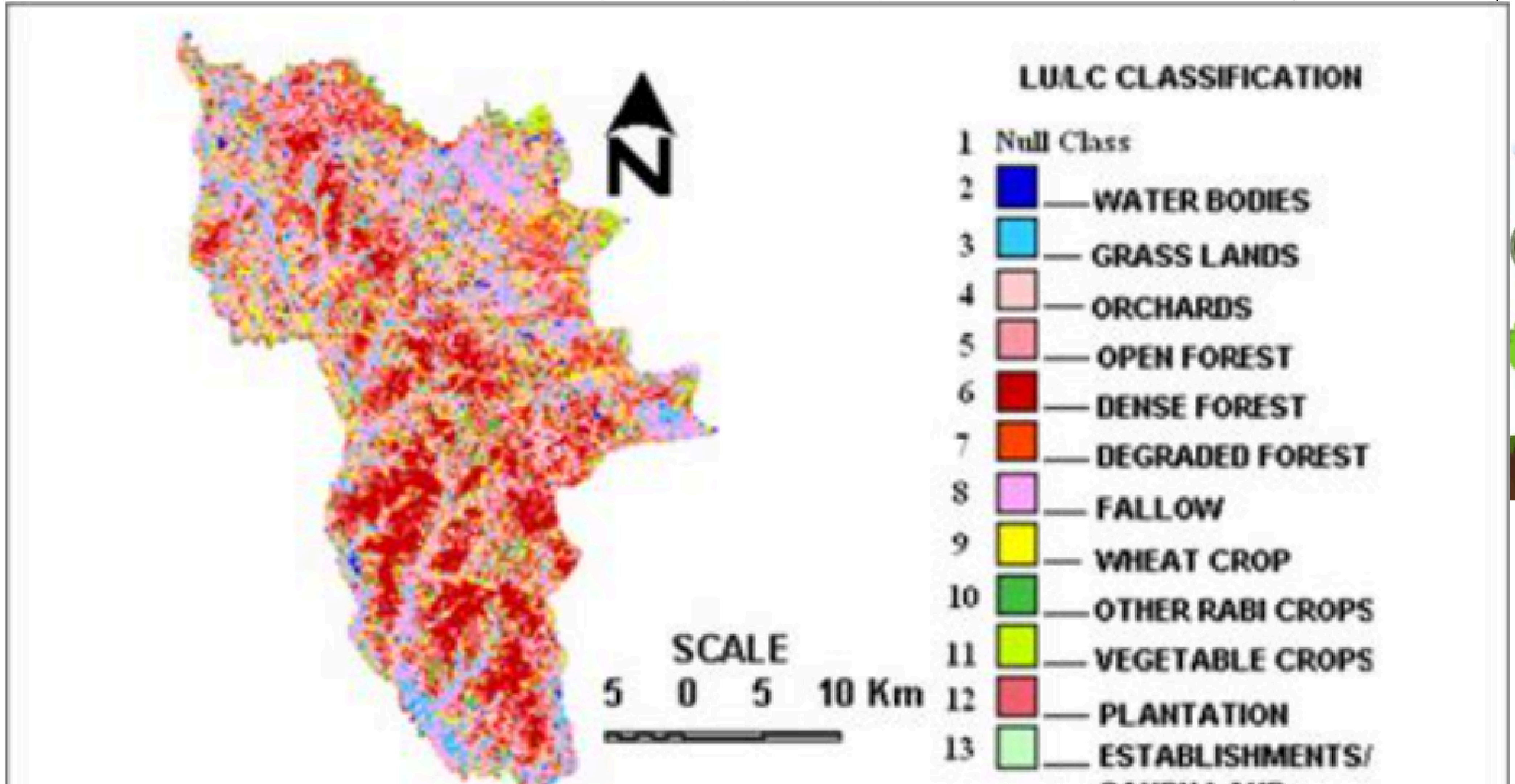
- **Performance** improved significantly with optimized **Spark configuration.**
- **ORC forma**t reduced **query time by ~45%** compared to default.
- For large **datasets** (>1 million records), **Spark performance** scaled **linearly.**
- Platform ensured **reproducibility** by packaging entire environments in **Docker containers**.
- Combined **waveform** and **EHR data analysis** was achieved.

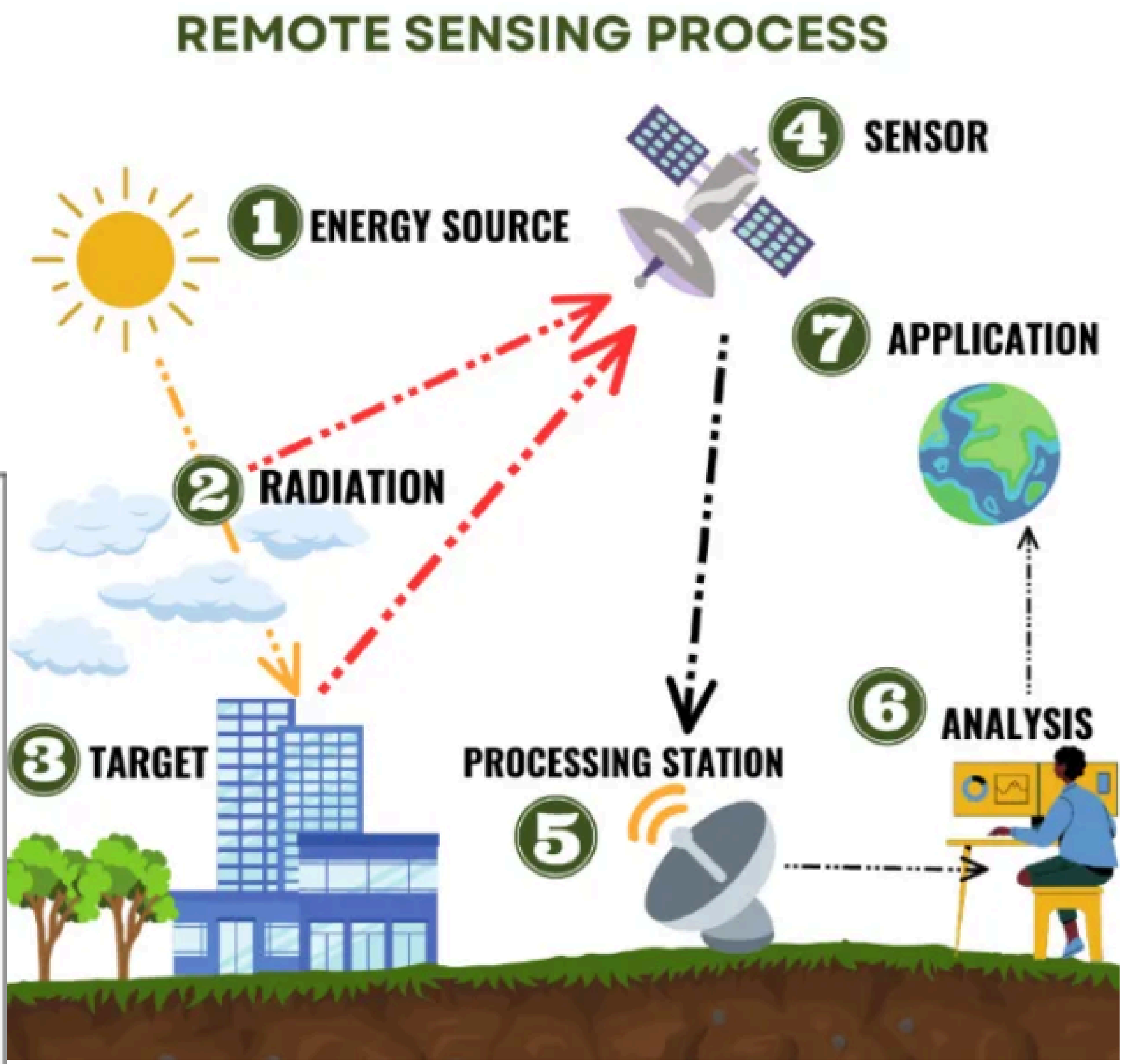
References

- Singh et al. (2018), Roy et al. (2015)
- NRSC satellite datasets
- Landsat TM, ETM+, OLI imagery
- ERDAS and ArcGIS documentation



A Energy Source of Illumination





Thank You