Minor Irrigation Census Analysis – Uttar Pradesh

- Big Data Analytics using Pyspark
- Real Time Demographic processing
- Policy-Driven Insights

Presented By: K .Vinay Balaji Reddy
Mentor : Dr. B. Jogeswara Rao

Introduction

- The Minor Irrigation Census provides crucial data on India’s irrigation resources and groundwater use.
- Objective: To analyze the **Uttar Pradesh dataset** (12,938 villages) to uncover irrigation efficiency and water stress trends.
- The project leverages **Apache Spark and PySpark** for distributed, large-scale data processing.
- Focus: Identifying patterns that support **data-driven agricultural and water management decisions**.

Project Objectives

Key Points:

- Process and analyze large-scale irrigation data using **Apache Spark**.
- Clean and validate 19 key water and irrigation metrics.
- Identify **district-wise irrigation efficiency** and **groundwater stress**.
- Generate **dynamic visual insights** using Python visualization libraries.
- Demonstrate **Big Data scalability** and performance improvements over traditional methods.

Technical Stack

Tools & Technologies Used:

- **Apache Spark 3.0+** – Distributed data processing framework
- **PySpark** – Python API for Apache Spark
- **Python 3.8+** – Programming language for data manipulation and analysis
- **Jupyter Notebook** – Interactive environment for running PySpark scripts and visualizations
- **Pandas, Matplotlib, Seaborn** – Libraries for data handling and visualization
- **Dataset:** *Minor Irrigation Census (Uttar Pradesh, 12,938 villages)*

Data Preprocessing & Analysis

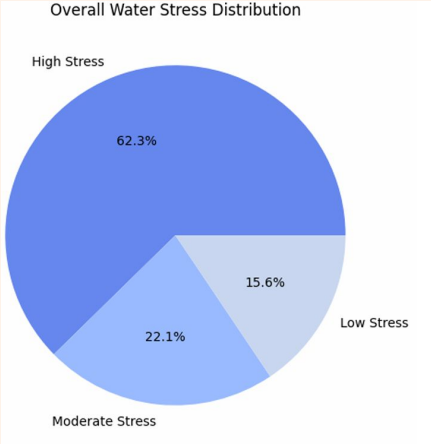
Content:

- **Data Cleaning:** Removed nulls, handled inconsistencies, validated key metrics.
- **Feature Selection:** Focused on 19 irrigation-related fields.
- **Distributed Processing:** Achieved up to **50× faster** computation with Spark.
- **Key Operations:**
    - GroupBy and Aggregations
    - Correlation analysis
    - Statistical summaries
    - Visualization preparation

Key Insights

Main Findings:

- 🌾 **Top-performing districts:** Aligarh, Etawah, Shahjahanpur, Unnao → 100% irrigation coverage
- 💧 **High-stress regions:** Maharajganj, Jaunpur → significant groundwater depletion
- 📏 **Average area per village:** 196 hectares
- 🔗 Strong correlation between **irrigation efficiency** and **sustainable groundwater levels**



Outcomes & Learnings

Results:

- Successful implementation of **Big Data workflow** using PySpark.
- Efficiently analyzed **~13K records** with accuracy and speed.
- Discovered **actionable irrigation insights** for Uttar Pradesh.

Skills Developed:

- Big Data processing with PySpark
- Data visualization and pattern recognition
- Data-driven decision-making and reporting

Conclusion

This Minor Irrigation Census Analysis Project leverages Apache Spark and PySpark to transform the way large-scale irrigation data is analyzed, demonstrating how distributed computing can efficiently handle **12,938 village records across Uttar Pradesh**. By systematically cleaning, validating, and analyzing 19 irrigation and water-related metrics, the project identifies **districts with full irrigation coverage** (Aligarh, Etawah, Shahjahanpur, Unnao) and **regions experiencing significant groundwater stress** (Maharajganj, Jaunpur).

The analysis reveals **average village area** patterns, uncovers correlations between **irrigation efficiency and sustainable groundwater levels**, and provides a quantitative foundation for prioritizing water resource management initiatives. Beyond technical achievement, the project exemplifies how **data-driven insights** can guide evidence-based agricultural planning and inform sustainable water governance at the district and state levels.

By combining scalable computation, accurate data processing, and actionable visualization, this project establishes a robust framework for **future irrigation and agricultural analytics**, representing a meaningful step toward leveraging big data for **resource optimization and informed decision-making** in India