DATA Warehouse

DSC 314 (Project)

Objective

- India’s agricultural economy is highly sensitive to climatic conditions such as rainfall, temperature, humidity, and extreme weather events. Rice,wheat and etc., being staple crops, play a crucial role in food security and international trade. Variations in climate directly affect crop yield, which in turn influences domestic availability, pricing, and decisions related to import and export.

- With the rapid growth of climate data, agricultural statistics, and trade records, traditional data analysis methods are no longer sufficient to extract meaningful insights. **Data Warehousing and Data Mining techniques** provide systematic approaches to store, integrate, analyze, and predict trends from large, heterogeneous datasets.

- This project focuses on building a **data warehouse integrating climate, crop production, and trade data**, and applying **data mining techniques** to predict climate patterns and their impact on rice and wheat import–export trends in India.

Key motivating factors include:

- Rapid increase in data volume
- Availability of low-cost storage and computing power
- Need for **knowledge discovery** rather than simple data retrieval
- Demand for **predictive and prescriptive analytics**

In the agricultural context, data mining is important because:

- Climate uncertainty affects crop yield
- Early prediction helps policymakers plan imports and exports
- Farmers and government agencies can minimize economic losses
- Long-term trends support sustainable agriculture planning

Rainfall Dataset Description and Data Collection

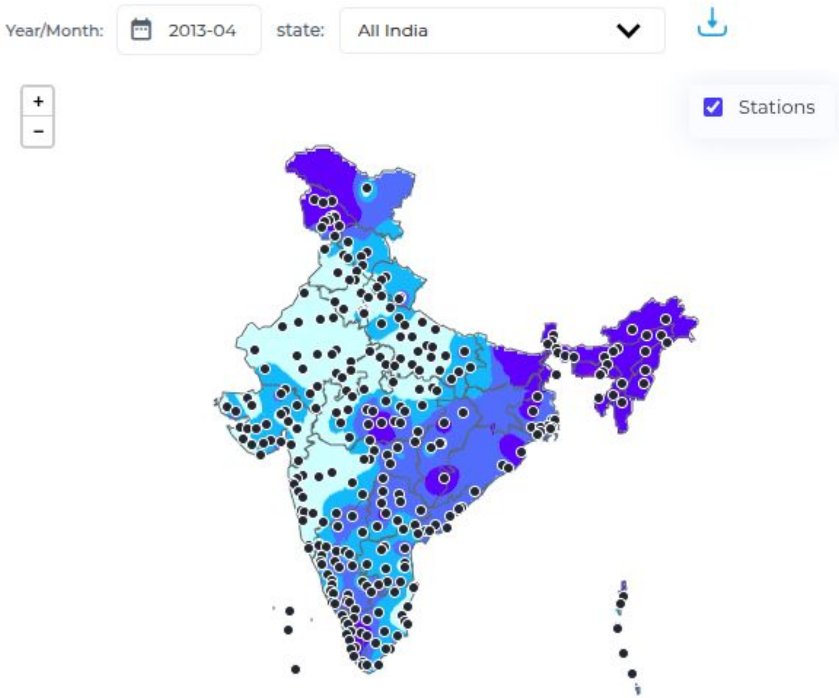
Source of the Dataset

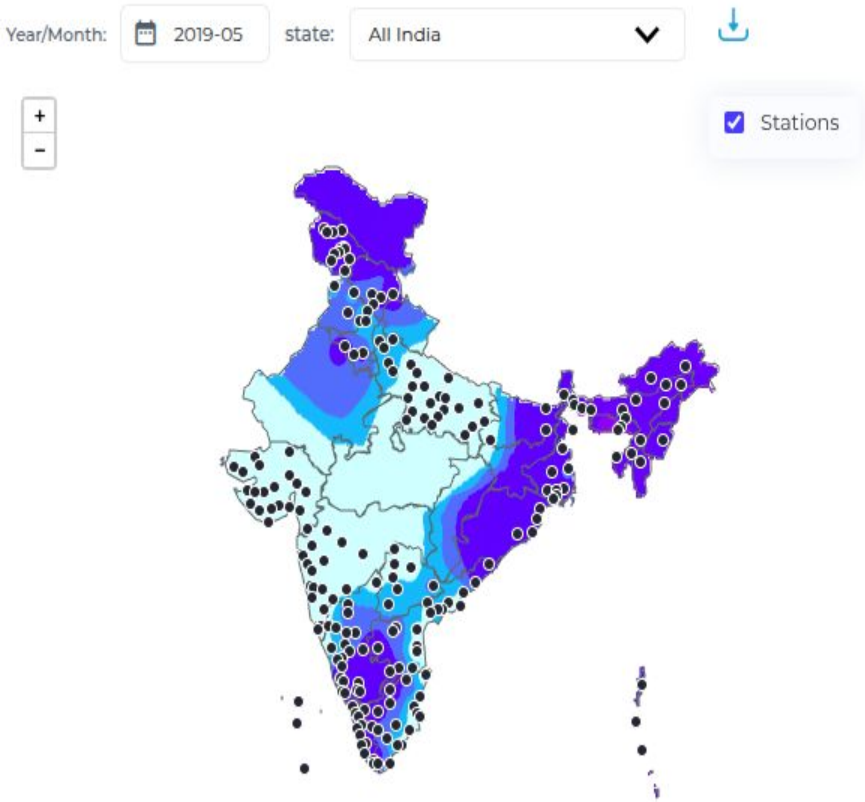
The rainfall dataset used in this project was obtained from the **India Climate & Energy Dashboard**, an official public data platform developed by national agencies in collaboration with research organizations. This dashboard provides authoritative and regularly updated climate and environmental data for India.

The platform aggregates rainfall observations from a wide network of meteorological stations distributed across all Indian states and union territories. Since the data is sourced from government and institutional monitoring systems, it is considered reliable and suitable for academic and policy-oriented analysis.

https://iced.niti.gov.in/climate-and-environment/climate-variability/rainfall

Station Name	State	District	Month	Year	Rainfall (mm)
PORT BLAIR	Andaman and Nicobar Islands	SOUTH ANDAMAN	Jan	2010	88.90
KHARGONE	Madhya Pradesh	WEST NIMAR	Jan	2010	0.00
UNA	Himachal Pradesh	UNA	Jan	2010	8.00
ADIRAMPATTINAM	Tamil Nadu	THANJAVUR	Jan	2010	22.90
ETAWAH	Uttar Pradesh	ETAWAH	Jan	2010	2.00
JODHPUR	Rajasthan	JODHPUR	Jan	2010	1.10
PURI	Odisha	PURI	Jan	2010	4.80
BANGALORE(A)	Karnataka	BENGALURU URBAN	Jan	2010	18.20
NAJIBABAD	Uttar Pradesh	BIJNORE	Jan	2010	5.00
NALGONDA	Telangana	NALGONDA	Jan	2010	36.00
HARDOI	Uttar Pradesh	HARDOI	Jan	2010	8.00
DURG	Chhattisgarh	DURG	Jan	2010	21.60
KHANDWA	Madhya Pradesh	EAST NIMAR	Jan	2010	0.00
TEHRI NEW	Uttarakhand	TEHRI NEW	Jan	2010	22.40
PARBHANI	Maharashtra	PARBHANI	Jan	2010	8.30
GAZIPUR	Uttar Pradesh	GAZIPUR	Jan	2010	3.00
VERAVAL	Gujarat	JUNAGAD	Jan	2010	0.00
TUNI	Andhra Pradesh	EAST GODAVARI	Jan	2010	2.80
PANJIM	Goa	GOA	Jan	2010	2.60
CHENNAI (MINAMBAKKAM (A))	Tamil Nadu	CHENNAI	Jan	2010	6.70
BAHRAICH	Uttar Pradesh	BAHRAICH	Jan	2010	0.00
PATIALA	Punjab	PATIALA	Jan	2010	12.40
BETUL	Madhya Pradesh	BETUL	Jan	2010	3.60
VELLORE	Tamil Nadu	VELLORE	Jan	2010	48.40
NORTH LAKHIMPUR(A) / LILABARI	Assam	LAKHIMPUR	Jan	2010	0.60
MUZAFFARNAGAR	Uttar Pradesh	MUZAFFAR NAGAR	Jan	2010	6.60
MORADABAD	Uttar Pradesh	MORADABAD	Jan	2010	0.40
DIAMOND HARBOUR	West Bengal	SOUTH 24 PARGANAS	Jan	2010	0.00
SURAT	Gujarat	SURAT	Jan	2010	2.80
JHALAWAR	Rajasthan	JHALAWAR	Jan	2010	5.00


Year/Month:
2013-04
state:
All India
Stations


Year/Month:
2019-05
state:
All India
Stations

Step 3: Data Cleaning, Formatting, and Final Export

Objective: Transform raw rainfall data into a clean master dataset suitable for analysis and storage.

Cleaning Logic Applied:

1. Garbage Removal

- Convert `Year` to numeric using coercion (`errors='coerce'`).
- Drop rows missing critical fields:
  - `Year`
  - `Month`
  - `Station Name`

2. Type Conversion

- Convert `Year` from float to integer.
- Convert `Rainfall (mm)` to numeric.

3. String Sanitization

- Remove leading and trailing whitespace from:
  - `State`
  - `District`
  - `Station Name`
  - `Month`

4. Directory Handling

- Create the `results` folder if it does not exist before saving output.

Output: `Final_Rainfall_Data_2010_2022.xlsx`

	A	B	C	D	E	F
1	Station Name	State	District	Month	Year	Rainfall (mm)
2	PILANI	Rajasthan	PILANI	Jan	2014	0
3	MANGALORE BAJPE(A)	Karnataka	DAKSHIN KANNADA	Jan	2014	0
4	JHALAWAR	Rajasthan	JHALAWAR	Jan	2014	29
5	TEZPUR	Assam	SONITPUR	Jan	2014	0.3
6	COIMBATORE / PEELAMEDU (A)	Tamil Nadu	COIMBATORE	Jan	2014	0
7	SURAT	Gujarat	SURAT	Jan	2014	13.4
8	JALPAIGURI	West Bengal	JALPAIGURI	Jan	2014	1.8
9	KOCHI A.P.(NEDUMBASSERY)	Kerala	ERNAKULAM	Jan	2014	0
10	SURENDRANAGAR	Gujarat	SURENDRANAGAR	Jan	2014	0
11	FURSATGANJ	Uttar Pradesh	RAIBARELI	Jan	2014	61.3
12	MADURAI(A)	Tamil Nadu	MADURAI	Jan	2014	10.2
13	AMINI DIVI	Lakshadweep	LAKSHADWEEP	Jan	2014	67.8
14	MUZAFFARNAGAR	Uttar Pradesh	MUZAFFAR NAGAR	Jan	2014	47.6
15	BERHAMPORE	West Bengal	MURSHIDABAD	Jan	2014	1
16	VELLORE	Tamil Nadu	VELLORE	Jan	2014	0.7

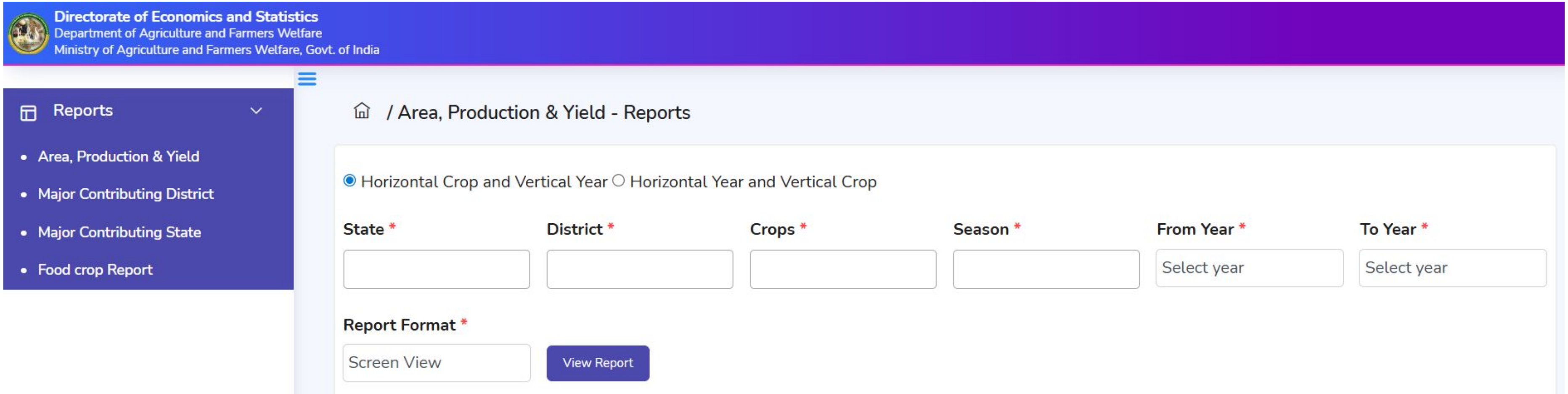
Crop Area, Production, and Yield Dataset Description

Source of the Dataset

The crop production dataset used in this project was obtained from the **Directorate of Economics and Statistics (DES)**, under the **Department of Agriculture and Farmers Welfare, Ministry of Agriculture and Farmers Welfare, Government of India**. The data is accessed through the official **Area, Production, and Yield (APY) Reports** portal.

This portal is an authoritative government source that provides comprehensive and officially validated agricultural statistics for India. The dataset is widely used for policy formulation, academic research, and economic analysis.

https://data.desagri.gov.in/website/crops-apy-report-web



State	District	Year	Banana			Coconut			Tapioca			Oilseeds total			Arecanut		
			Whole Year			Whole Year			Whole Year			Whole Year			Whole Year		
			Area (Hectare)	Production (Tonnes)	Yield (Tonne/Hectare)	Area (Hectare	oduction (Nu	d (Nuts/Hect	Area (Hectare	duction (Ton	l (Tonne/Hec	Area (Hectare	duction (Ton	l (Tonne/Hec	Area (Hectare	duction (Ton	l (Tonne/Hec
		2010 - 2011	593	3425	5.78	14560	71300000	4896.98	69	575	8.33						
		2011 - 2012	620.5	3720	6	14590	69700000	4777.24	64	505	7.89						
		2012 - 2013	241	2034	8.44	14650	89800000	6129.69	61	523	8.57						
		2013 - 2014	170	1300.5	7.65	14655	96200000	6564.31	22	195.3	8.88						
		2014 - 2015	170.5	1570	9.21	14673	95300000	6494.92	46	527.3	11.46						
		2015 - 2016	517	1920	3.71	14676	94500000	6439.08	54.5	745	13.67						
		2016 - 2017	679.5	1495	2.2	14840	96200000	6482.48	72	297	4.13						
		2017 - 2018	790	46	0.06	13695	98130000	7165.39	48.3	416.1	8.61						
		2018 - 2019	803.57	82.78	0.1	13728	70360000	5125.29	45.11	304.42	6.75						
		2019 - 2020	932.38	609.06	0.65	15728	89060000	5662.51	61.11	474.42	7.76						
	1. Nicobars	2020 - 2021	409.83	752.97	1.84	13748	99859000	7263.53									
		2010 - 2011	657	7968	12.13	3668	12700000	3462.38	186.5	1355	7.27						
		2011 - 2012	682	9085	13.32	3668	15600000	4253	181.5	1280.5	7.06	71.5	38.5	0.54			
		2012 - 2013	1055.5	10588.5	10.03	3675	17800000	4843.54	190	1446	7.61	93.4	28.3	0.3			
		2013 - 2014	1343.5	9263.8	6.9	3685	16800000	4559.02	184.2	3503.8	19.02	39.2	13.31	0.34			
		2014 - 2015	1352	10790	7.98	3675	17400000	4734.69	137.8	2161.9	15.69	34.4	13.8	0.4			
		2015 - 2016	1178	11180	9.49	3675	19500000	5306.12	26.5	600	22.64						
		2016 - 2017	1012	5226	5.16	3675	20000000	5442.18	41	1010	24.63						
		2017 - 2018	584.85	4944.09	8.45	1347.3	10090000	7489.05	40.9	694.04	16.97						
		2018 - 2019	650.54	6999.44	10.76	1525.63	9490000	6220.38	32.58	383.85	11.78						
		2019 - 2020	665.12	5904.3	8.88	1532.48	7610000	4965.81	29.13	213.58	7.33						
	2. North and m	2020 - 2021	935.98	7310	7.81	1537.38	6560000	4267									
		2010 - 2011	360	5517	15.33	3540	11000000	3107.34	22.5	220	9.78						
		2011 - 2012	378.5	5730	15.14	3542	19700000	5561.83	19.5	260	13.33	20	8	0.4			
		2012 - 2013	378.5	5727.5	15.13	3550	17500000	4929.58	19	151	7.95	15.4	23.6	1.53			
		2013 - 2014	304	3478	11.44	3560	16000000	4494.38	33	547.5	16.59	1.2	0.92	0.77			
		2014 - 2015	318.5	3602	11.31	3562	17100000	4800.67	28.5	575.8	20.2	1.6	1.2	0.75			
		2015 - 2016	320	4550	14.22	3564	17600000	4938.27	26.5	725	27.36						
		2016 - 2017	444.6	6623.3	14.9	3564	16800000	4713.8	20	762	38.1						
		2017 - 2018	440	8622.4	19.6	1232.5	16560000	13436.11	24	245.4	10.23						
		2018 - 2019	448	9607.9	21.45	2895	16330000	5640.76	12.49	637.31	51.03						
		2019 - 2020	438.7	9602.06	21.89	2810	15880000	5651.25	12.24	473.05	38.65						
1. Andaman ar	3. South andan	2020 - 2021	410.24	8260.3	20.14	2810	18950000	6743.77									

Step 7: ETL Pipeline — Cleaning and Reshaping Crop Data

Objective: Transform wide-format crop data into a normalized long-format dataset.

Transformation Steps:

1. Header Parsing

Extract metric names (`Area`, `Production`, `Yield`) embedded within the first data row.

2. Identifier Cleaning

- Remove numbering prefixes from `State` and `District`.
- Convert year ranges into integer years.

3. Reshaping (Wide to Long)

Convert crop-specific columns into rows so each record represents:

- `State`
- `District`
- `Year`
- `Crop`
- `Area`
- `Production`
- `Yield`

	A	B	C	D	E	F	G
1	State	District	Year	Area	Production	Yield	Crop
2	Andhra Pradesh	Anantapur	2010	21	236	11.24	Potato
3	Andhra Pradesh	Anantapur	2011	18	181	10.06	Potato
4	Andhra Pradesh	Chittoor	2010	1138	10817	9.51	Potato
5	Andhra Pradesh	Chittoor	2011	1151	15206	13.21	Potato
6	Andhra Pradesh	Kadapa	2010	1	11	11	Potato
7	Andhra Pradesh	Kurnool	2010	14	157	11.21	Potato
8	Andhra Pradesh	Kurnool	2011	32	321	10.03	Potato
9	Andhra Pradesh	Visakhapatanam	2010	96	1076	11.21	Potato
10	Andhra Pradesh	Visakhapatanam	2011	46	462	10.04	Potato
11	Andhra Pradesh	Vizianagaram	2010	2	22	11	Potato
12	Arunachal Pradesh	Anjaw	2010	110	935	8.5	Potato
13	Arunachal Pradesh	Anjaw	2011	117	1043	8.91	Potato
14	Arunachal Pradesh	Anjaw	2012	120	1022	8.52	Potato
15	Arunachal Pradesh	Anjaw	2013	120	1022	8.52	Potato
16	Arunachal Pradesh	Anjaw	2014	33	138	4.18	Potato
17	Arunachal Pradesh	Anjaw	2015	73	392	5.37	Potato
18	Arunachal Pradesh	Anjaw	2016	73	392	5.37	Potato

Output: `Final_Crop_Data_2010_2022.xlsx`

Part 3: Data Integration and Merging

Step 8: Gap Analysis and Manual Mapping

Objective: Identify district mismatches between rainfall and crop datasets.

Examples of Issues:

- "SPSR Nellore" vs "Nellore"
- Minor spelling variations
- Case sensitivity differences

Process:

1. Convert names to uppercase for uniform comparison.
2. Compare crop districts against valid rainfall districts.
3. Generate:
    - `reference_district_list.txt`
    - `manual_mapping_worksheet.csv`

These files allow manual correction of mismatched district names.

Part 4: Final Integration & Engineering

Step 9: Applying Manual Geographic Mapping

Objective: Resolve mismatched District names between the Agriculture and Climate datasets.

Methodology:

Instead of relying on fuzzy matching (which produced low accuracy), we apply a hardcoded dictionary containing 333 manual corrections.

- **Source:** Manual audit of the Mismatch Report
- **Logic:** Maps district name variations such as `VISAKHAPATANAM` → `VISAKHAPATNAM`
- **Execution:** A new column `District_Final` is created and used as the joining key.

Output: `Final_Merged_Dataset_Clean.xlsx` (Preliminary Merge)

Step 10: Final Engineering (Aggregation & Validation)

Objective: Perform final cleanup logic to prepare the dataset for modeling and database storage.

Processing Steps:

1. Duplicate Handling

Aggregate duplicate seasonal entries (e.g., Kharif and Rabi) by summing:

- `Area`
- `Production`

2. Case-Sensitivity Fix

Ensure perfect joins by creating uppercase matching keys for `State`.

3. Missing Data Removal

Drop rows with `NaN` rainfall values to create an ML-ready dataset.

Mathematical Logic

Final Yield is computed as:

$$Yield_{Final} = \frac{\sum Production}{\sum Area}$$

Output: `Final_Engineered_Dataset.csv` (12,426 rows)

Step 11: Building the Final OLTP Database (Normalization)

Objective: Store the clean, merged dataset into a relational SQLite database using 3rd Normal Form (3NF) to minimize redundancy.

For example, long strings such as `"Andaman and Nicobar Islands"` are stored once and referenced using IDs.

Schema Design

Dimension Tables

- **States**
  - `StateID`
  - `StateName`
- **Districts**
  - `DistrictID`
  - `DistrictName`
  - `StateID` (Foreign Key)
- **Crops**
  - `CropID`
  - `CropName`

Fact Table

- **Crop_Yield_Facts**
  - `FactID`
  - `Year`
  - `Area`
  - `Production`
  - `Yield`
  - `Rainfall`
  - `DistrictID` (Foreign Key)
  - `CropID` (Foreign Key)

Output: `Final_Agri_Weather_OLTP.db`

Step 12: Final OLAP Analysis (Business Intelligence)

Objective: Perform multi-dimensional analysis on the engineered dataset to extract meaningful insights.

We simulate an OLAP Cube using Pandas.

Operations Performed

1. Roll-Up

Aggregate production by `State` to identify the highest producing regions.

2. Dice

Filter for a specific sub-cube, for example:

- Crop = Rice
- High rainfall years

3. Slice

Isolate a specific year (e.g., 2014) to compare crop performance.

4. Pivot

Create a cross-tabulation of Yield trends over the years.

5. Correlation Analysis

Analyze whether rainfall has a measurable impact on yield.

Example question: Does increased rainfall significantly increase crop productivity?

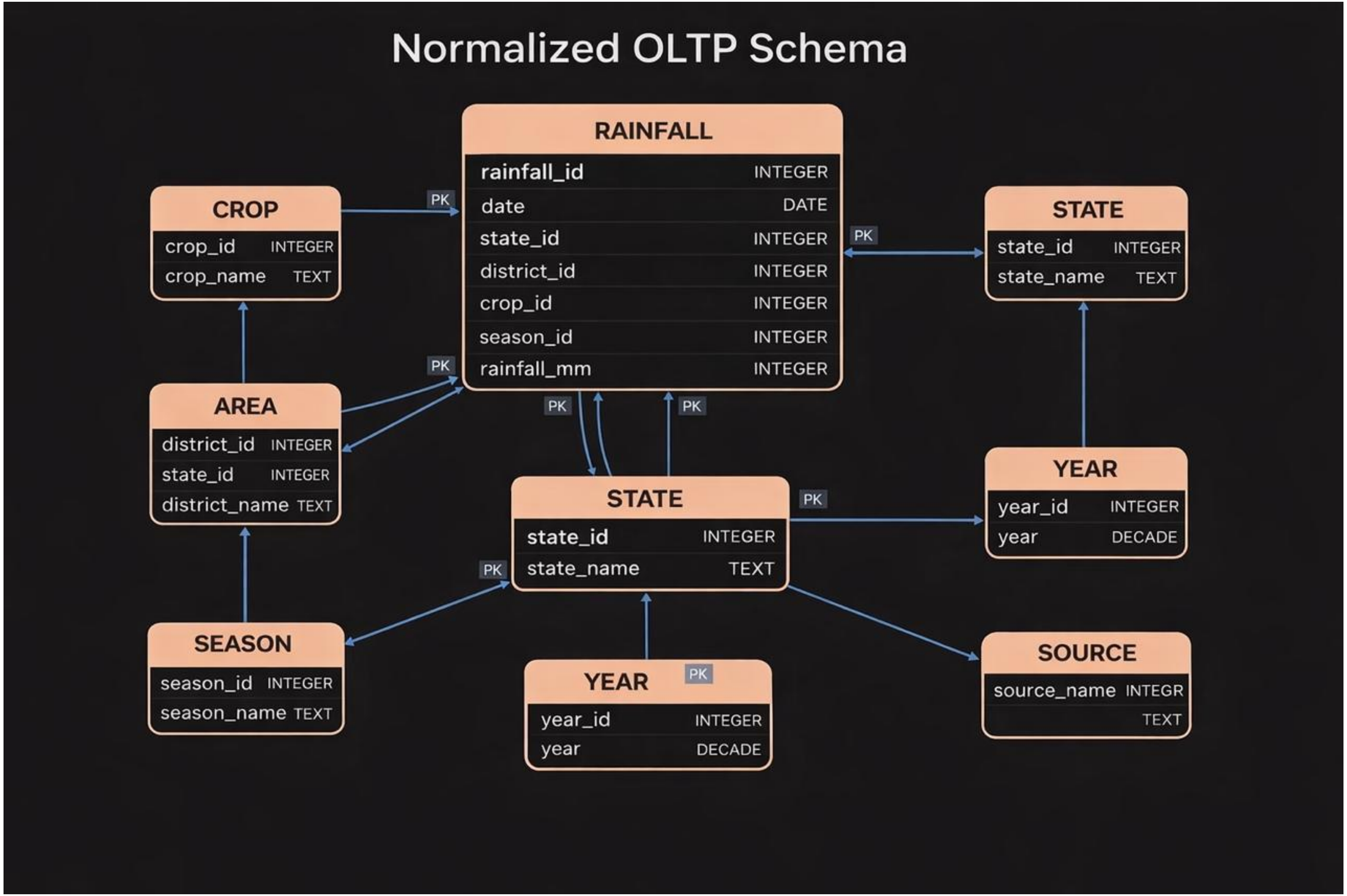
Final Outcome: A fully engineered Agriculture + Climate dataset stored in a normalized SQL database and analyzed using OLAP-style multi-dimensional operations.

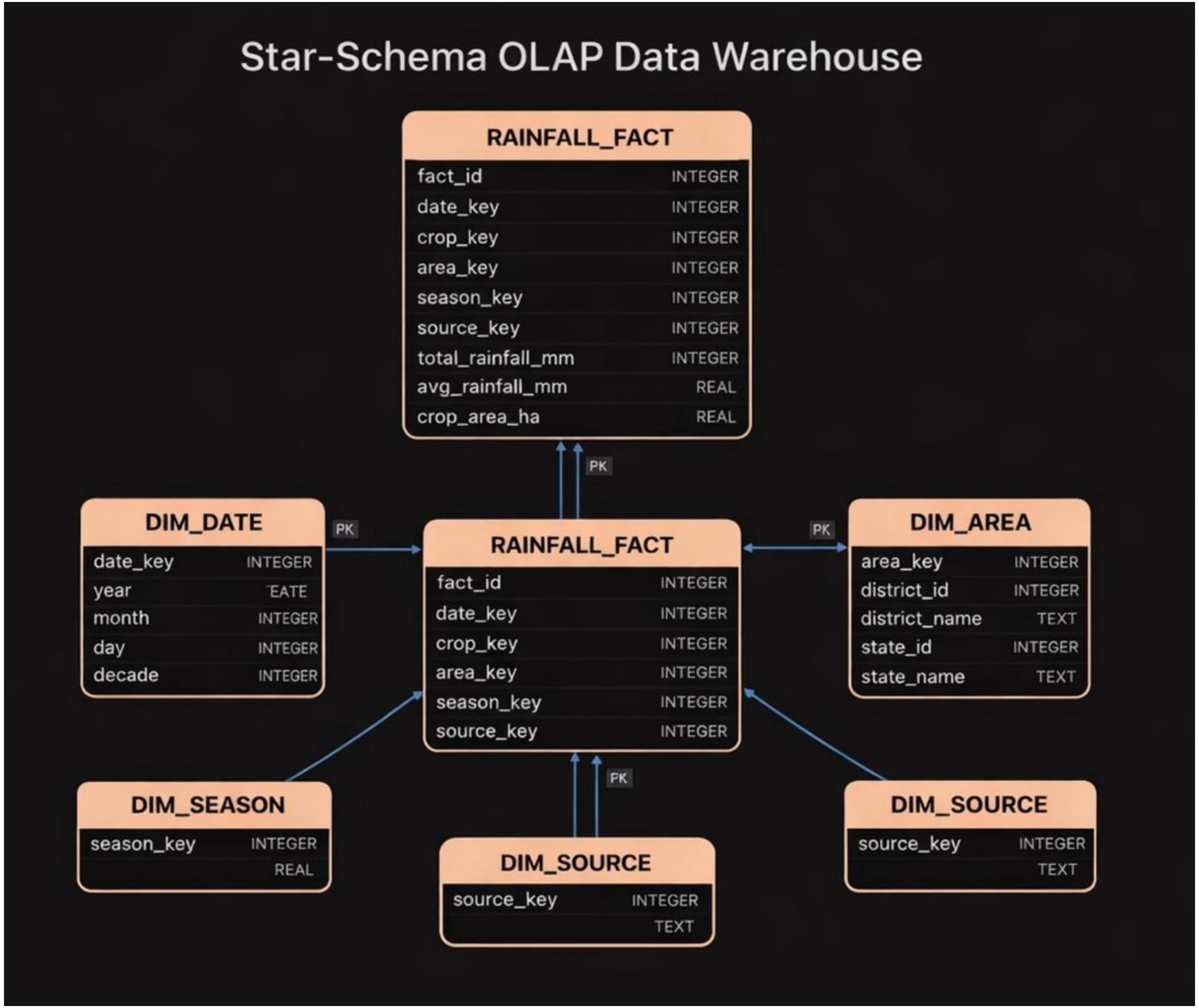
```python
df.head()
```

[13] 0.0s

	State	District	Year	Crop	Area	Production	Yield	Annual_Rainfall
0	Andaman and Nicobar Islands	NICOBAR	2012	Banana	241.0	2034.0	8.439834	5815.4
1	Andaman and Nicobar Islands	NICOBAR	2012	Coconut	14650.0	89800000.0	6129.692833	5815.4
2	Andaman and Nicobar Islands	NICOBAR	2012	Tapioca	61.0	523.0	8.573770	5815.4
3	Andaman and Nicobar Islands	NICOBAR	2013	Banana	170.0	1300.5	7.650000	6011.0
4	Andaman and Nicobar Islands	NICOBAR	2013	Coconut	14655.0	96200000.0	6564.312521	6011.0

Generate + Code + Markdown


Normalized OLTP Schema
RAINFALL
rainfall_id INTEGER
date DATE
state_id INTEGER
district_id INTEGER
crop_id INTEGER
season_id INTEGER
rainfall_mm INTEGER
CROP
crop_id INTEGER
crop_name TEXT
STATE
state_id INTEGER
state_name TEXT
AREA
district_id INTEGER
state_id INTEGER
district_name TEXT
STATE
state_id INTEGER
state_name TEXT
YEAR
year_id INTEGER
year DECADE
SEASON
season_id INTEGER
season_name TEXT
YEAR
year_id INTEGER
year DECADE
SOURCE
source_name INTEGR
TEXT
PK



Star-Schema OLAP Data Warehouse
RAINFALL_FACT
fact_id INTEGER
date_key INTEGER
crop_key INTEGER
area_key INTEGER
season_key INTEGER
source_key INTEGER
total_rainfall_mm INTEGER
avg_rainfall_mm REAL
crop_area_ha REAL
PK
DIM_DATE
date_key INTEGER
year EATE
month INTEGER
day INTEGER
decade INTEGER
PK
RAINFALL_FACT
fact_id INTEGER
date_key INTEGER
crop_key INTEGER
area_key INTEGER
season_key INTEGER
source_key INTEGER
PK
DIM_AREA
area_key INTEGER
district_id INTEGER
district_name TEXT
state_id INTEGER
state_name TEXT
FK
DIM_SEASON
season_key INTEGER
REAL
DIM_SOURCE
source_key INTEGER
TEXT
DIM_SOURCE
source_key INTEGER
TEXT