Agricultural Crop Production & Yield Optimization Analytics System

End-to-End Data Engineering & Analytics Capstone

Agneepradeep Verma

PROJECT OVERVIEW

Automated Analytics Pipeline for Agricultural Intelligence

Automated Pipeline

End-to-end agricultural analytics pipeline with minimal manual intervention

Medallion Architecture

Bronze, Silver, Gold data layers with Airflow orchestration for production-grade reliability

Interactive BI Dashboards

Power BI dashboards enabling real-time insights and data-driven decision making

Technology Stack: Python | Pandas | PySpark | Databricks | Apache Airflow | Power BI

Business Context & Problem Statement

Core Challenges

- **Data Volume & Distribution:** Agricultural data is large-scale, geographically distributed, and spans multiple dimensions (crops, regions, seasons, time)
- **Manual Processing Limitations:** Traditional manual analysis methods create bottlenecks and prevent timely, actionable insights
- **Scalability Requirements:** Growing data volumes demand automated, scalable analytics infrastructure

Solution Focus: Production-grade data pipeline enabling automated, enterprise-level agricultural analytics

DATA ARCHITECTURE

Medallion Architecture Implementation

Bronze Layer: Raw Ingestion

- Python & Pandas for initial data loading
- Schema standardization across source files
- File-based storage with metadata tracking

Silver Layer: Transformation & Cleansing

- Data cleaning and quality validation
- Year parsing and temporal alignment
- Yield recalculation and unit normalization
- Analysis-ready normalized dataset

Gold Layer: Star Schema Modeling

- Fact and dimension table creation
- Star schema optimization for BI tools
- Performance-tuned aggregations

Airflow Workflow Orchestration

Docker-Based Orchestration

Apache Airflow provides robust pipeline management with clear dependency handling and modular task design.

01 ingest_raw_to_bronze

Raw data ingestion task

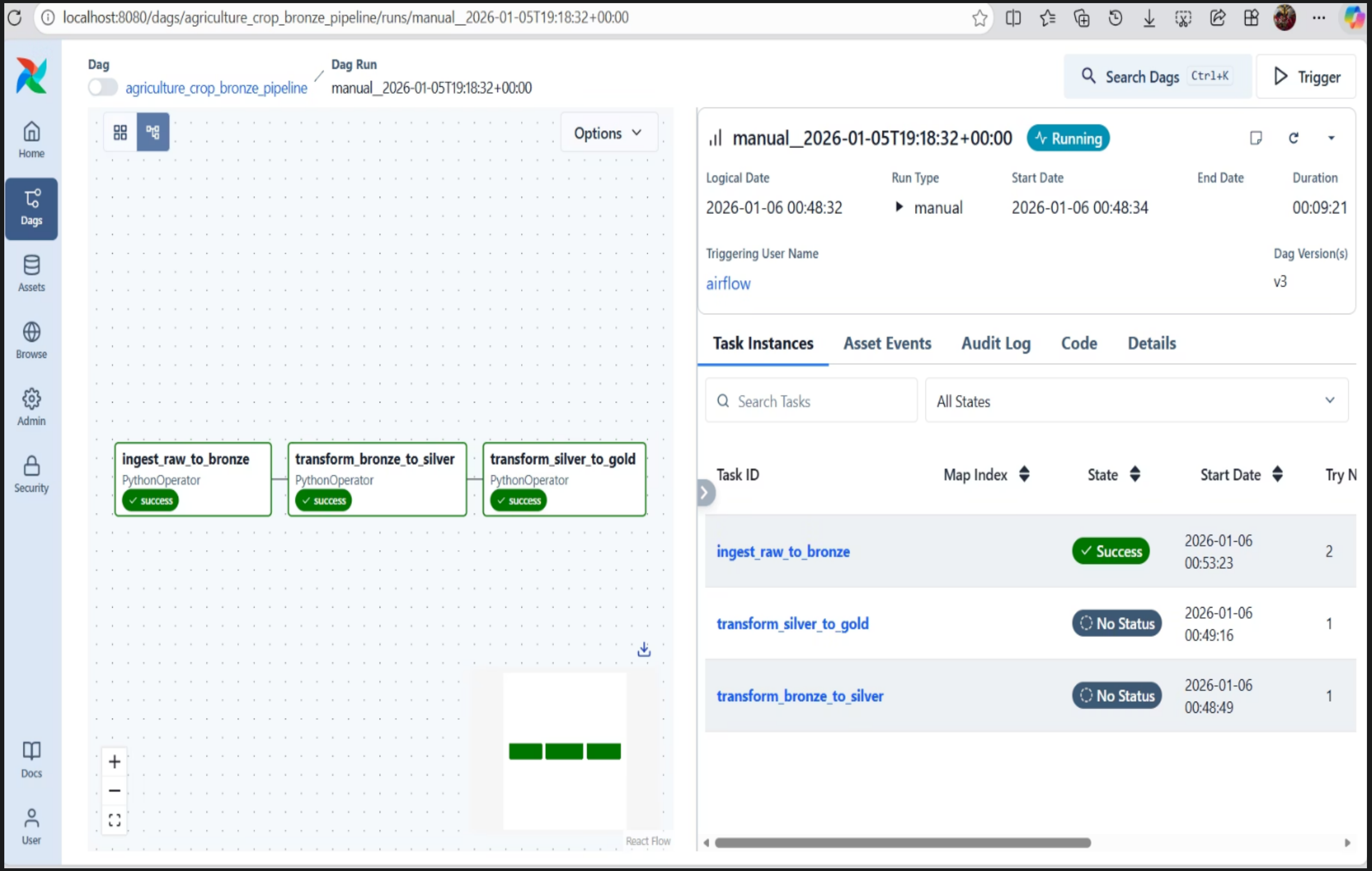
02 transform_bronze_to_silver

Data cleaning and validation

03 transform_silver_to_gold

Star schema generation

Manual trigger DAG enables controlled execution and testing



localhost:8080/dags/agriculture_crop_bronze_pipeline/runs/manual__2026-01-05T19:18:32+00:00

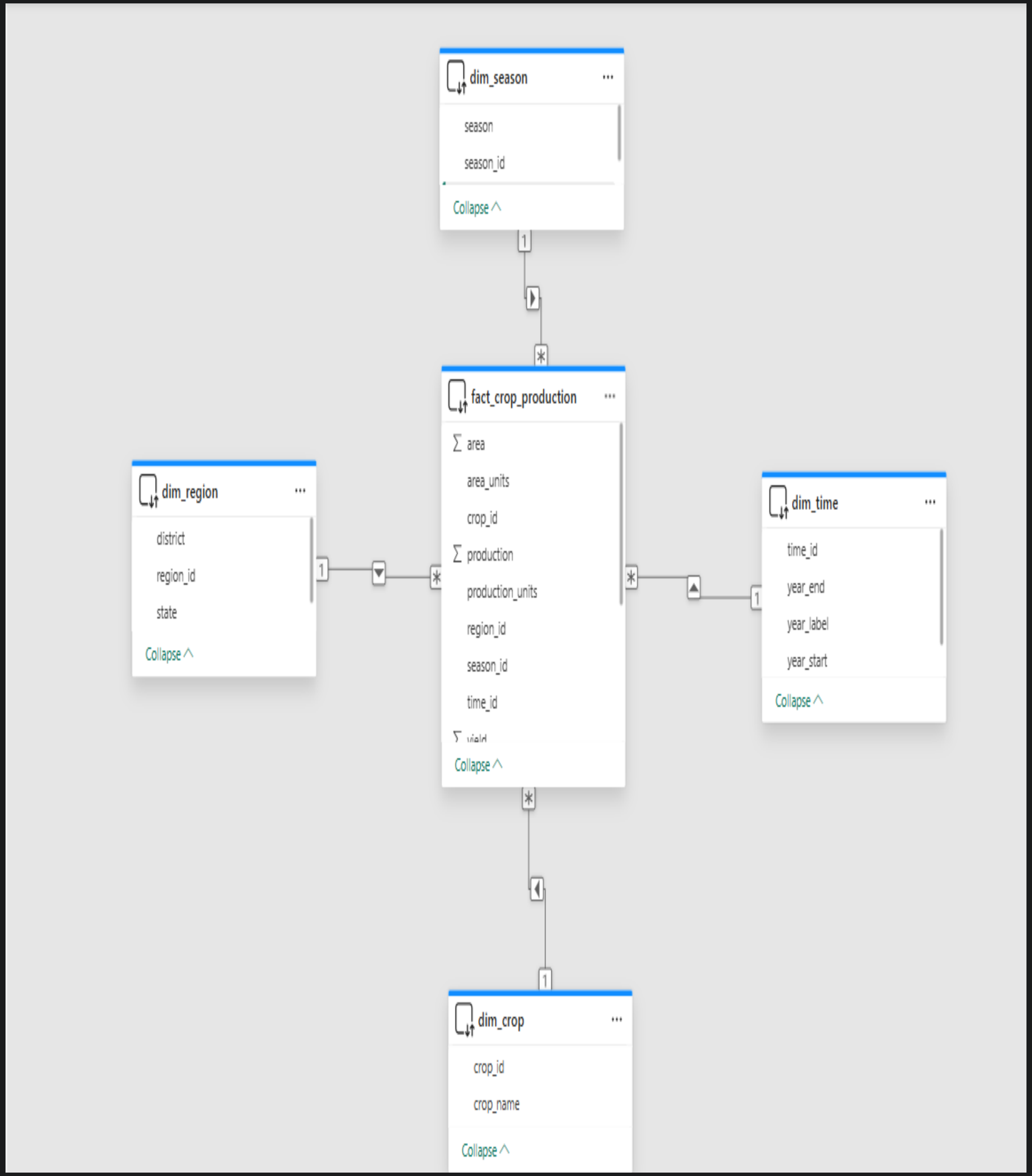
Dag: agriculture_crop_bronze_pipeline / Dag Run: manual__2026-01-05T19:18:32+00:00

manual__2026-01-05T19:18:32+00:00 Running

Logical Date: 2026-01-06 00:48:32
Run Type: manual
Start Date: 2026-01-06 00:48:34
End Date:
Duration: 00:09:21
Triggering User Name: airflow
Dag Version(s): v3

Task ID	Map Index	State	Start Date	Try N
ingest_raw_to_bronze		✓ Success	2026-01-06 00:53:23	2
transform_silver_to_gold		No Status	2026-01-06 00:49:16	1
transform_bronze_to_silver		No Status	2026-01-06 00:48:49	1

Data Modeling: Gold Layer Star Schema



Dimension Tables

- **Crop:** Crop types and categories
- **Region:** State and district hierarchy
- **Season:** Agricultural season definitions
- **Time:** DateTime dimensions

Why Star Schema?

- Simplifies complex analytics queries
- Optimizes Power BI performance
- Enables intuitive slicing, filtering, and drill-through navigation

Central Fact Table: Crop Production metrics connecting all dimensions for comprehensive analysis

Power BI: Executive Overview Dashboard

High-Level KPIs

Total Production, Cultivated Area, and Average Yield metrics

Interactive Controls

Year slicer for temporal analysis

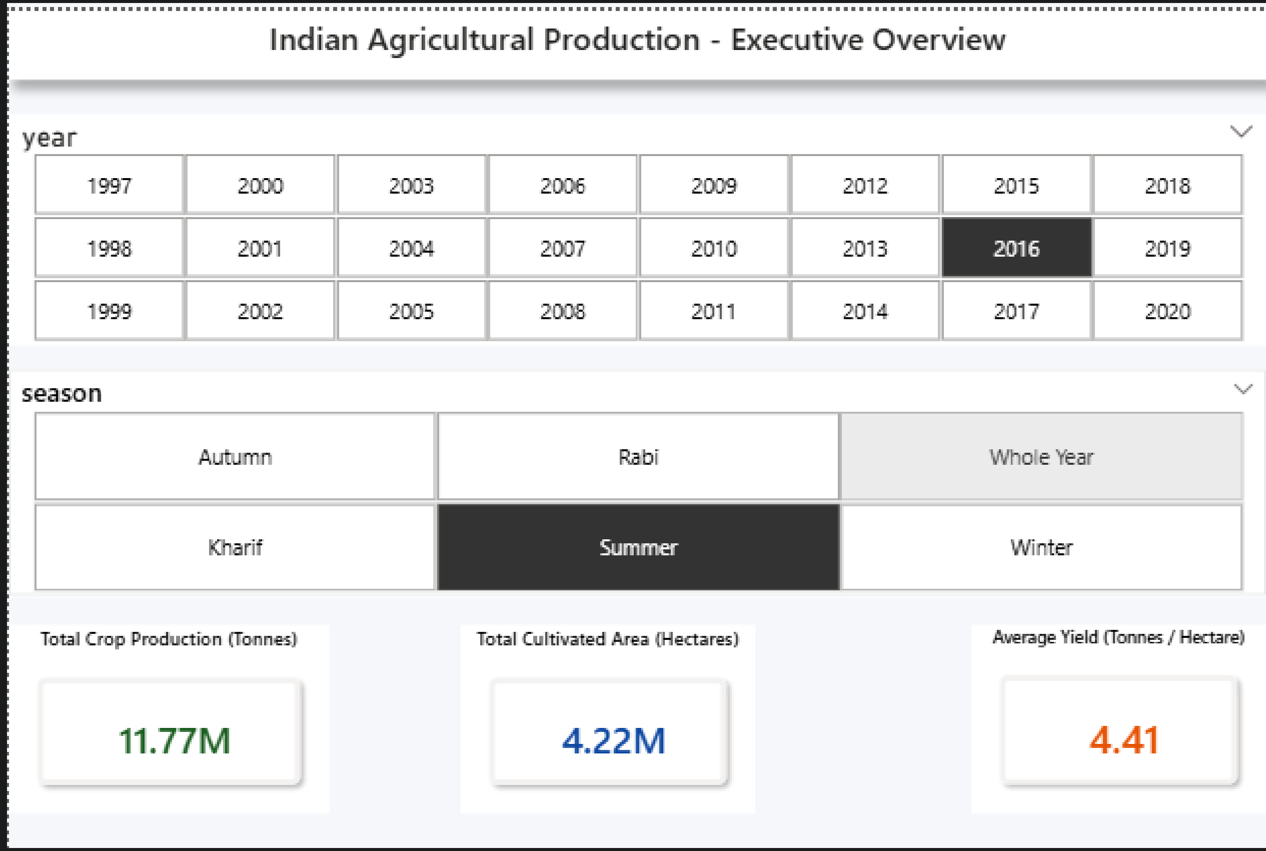
Top Performers

Leading crops by production volume

Geographic Visualization

State-level production heat map

Technical Integration: Connected directly to Gold layer with optimized DAX measures for real-time performance



Indian Agricultural Production - Executive Overview

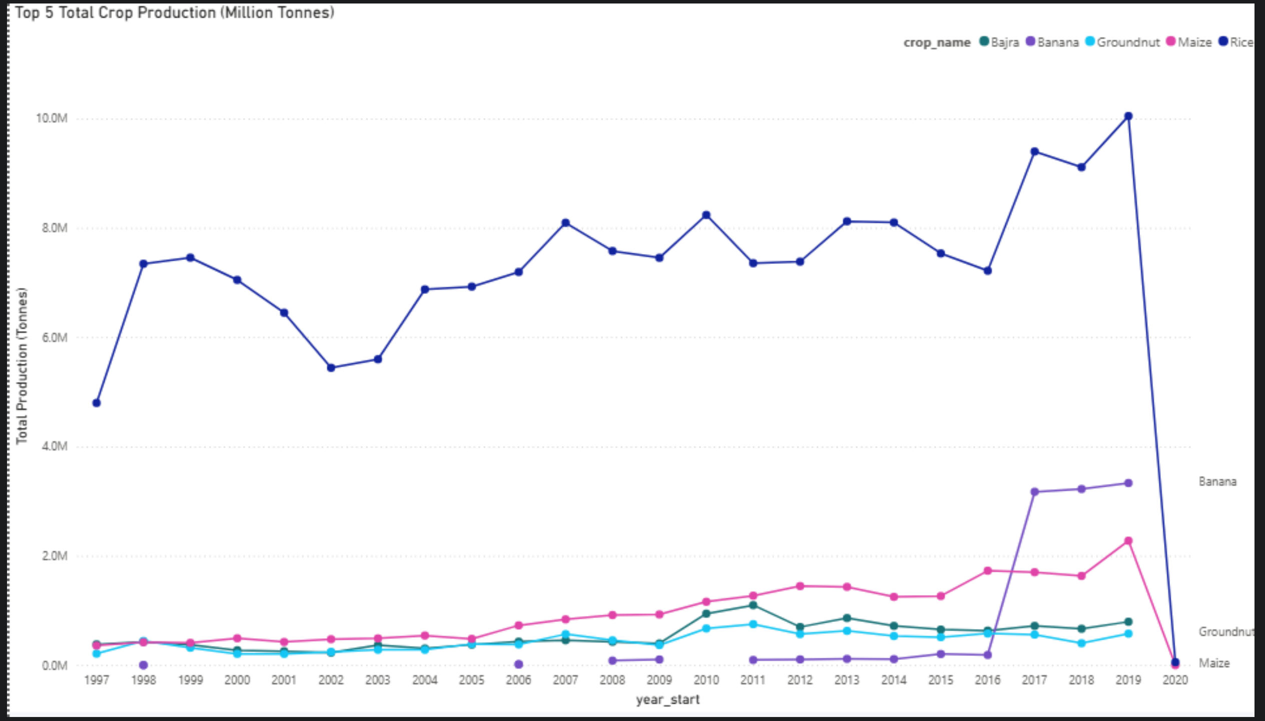
year

1997	2000	2003	2006	2009	2012	2015	2018
1998	2001	2004	2007	2010	2013	[✓ 2016]	2019
1999	2002	2005	2008	2011	2014	2017	2020

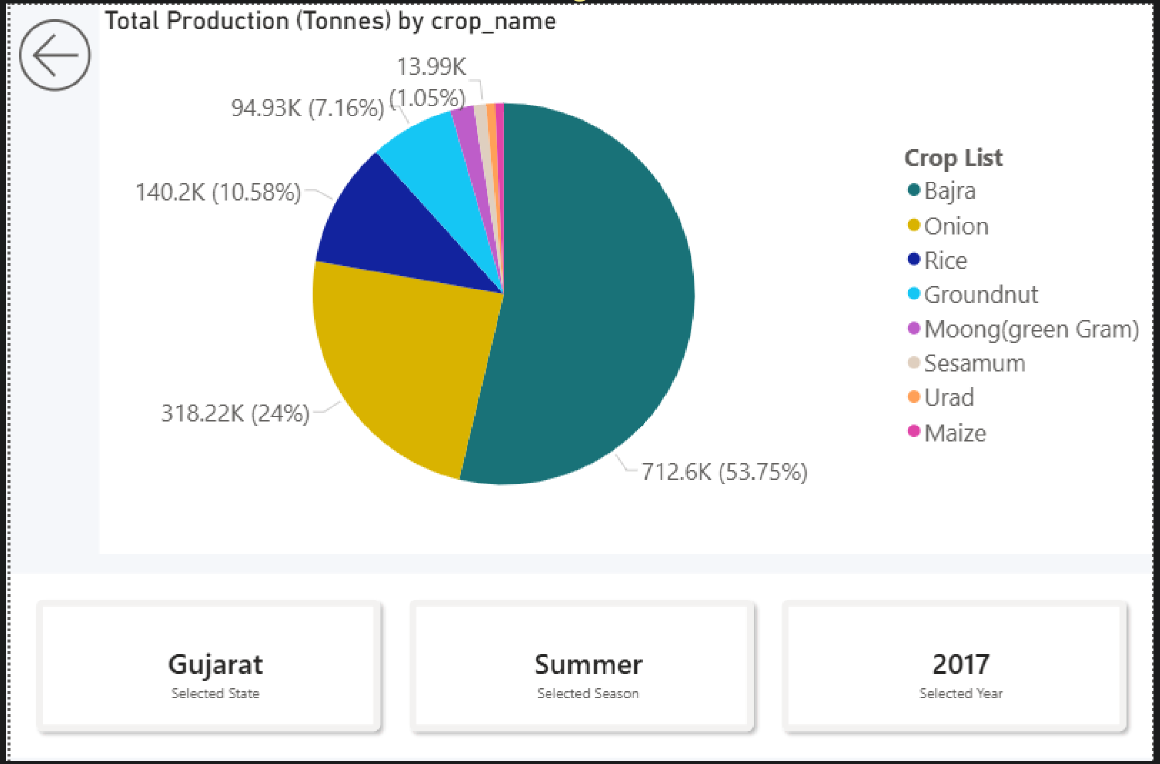
season

Autumn	Rabi	Whole Year
Kharif	[✓ Summer]	Winter

Total Crop Production (Tonnes)	Total Cultivated Area (Hectares)	Average Yield (Tonnes / Hectare)
11.77M	4.22M	4.41



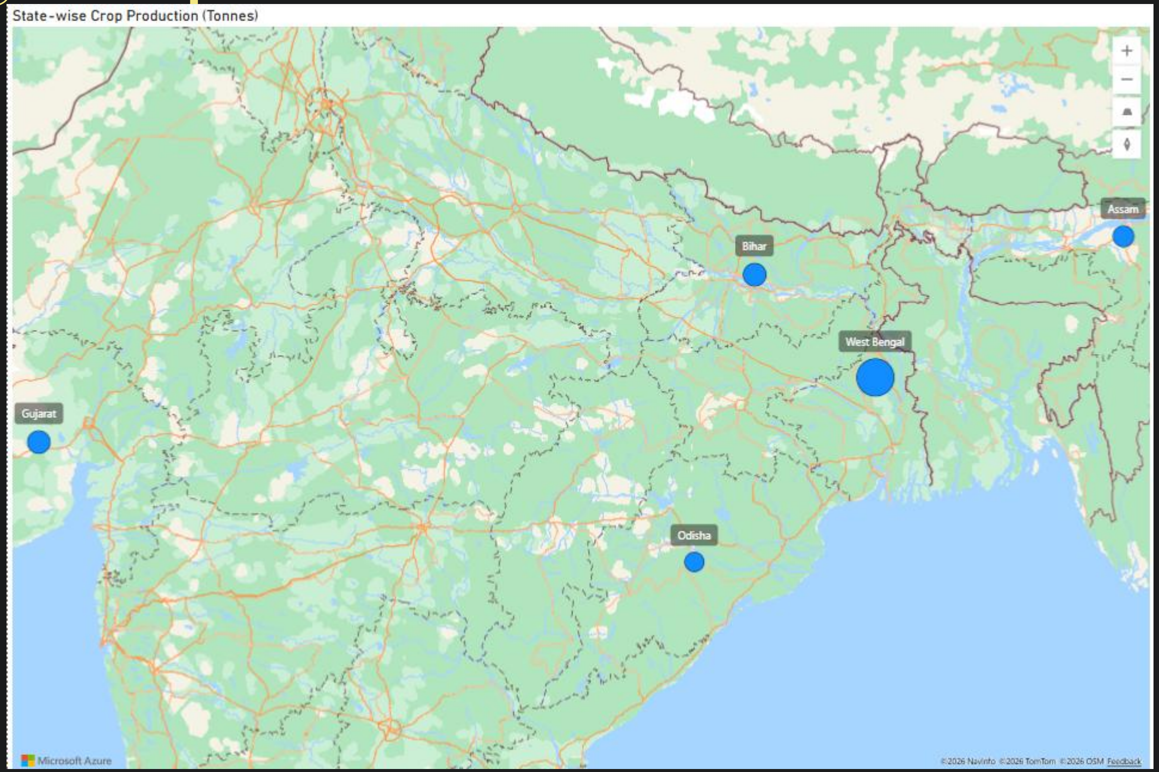
Advanced Analytics & Drill-Through Capabilities



Gujarat
Selected State

Summer
Selected Season

2017
Selected Year


State-wise Crop Production (Tonnes)
Gujarat
Bihar
Assam
West Bengal
Odisha
Microsoft Azure
©2026 NavInfo ©2026 TomTom ©2026 OSM Feedback

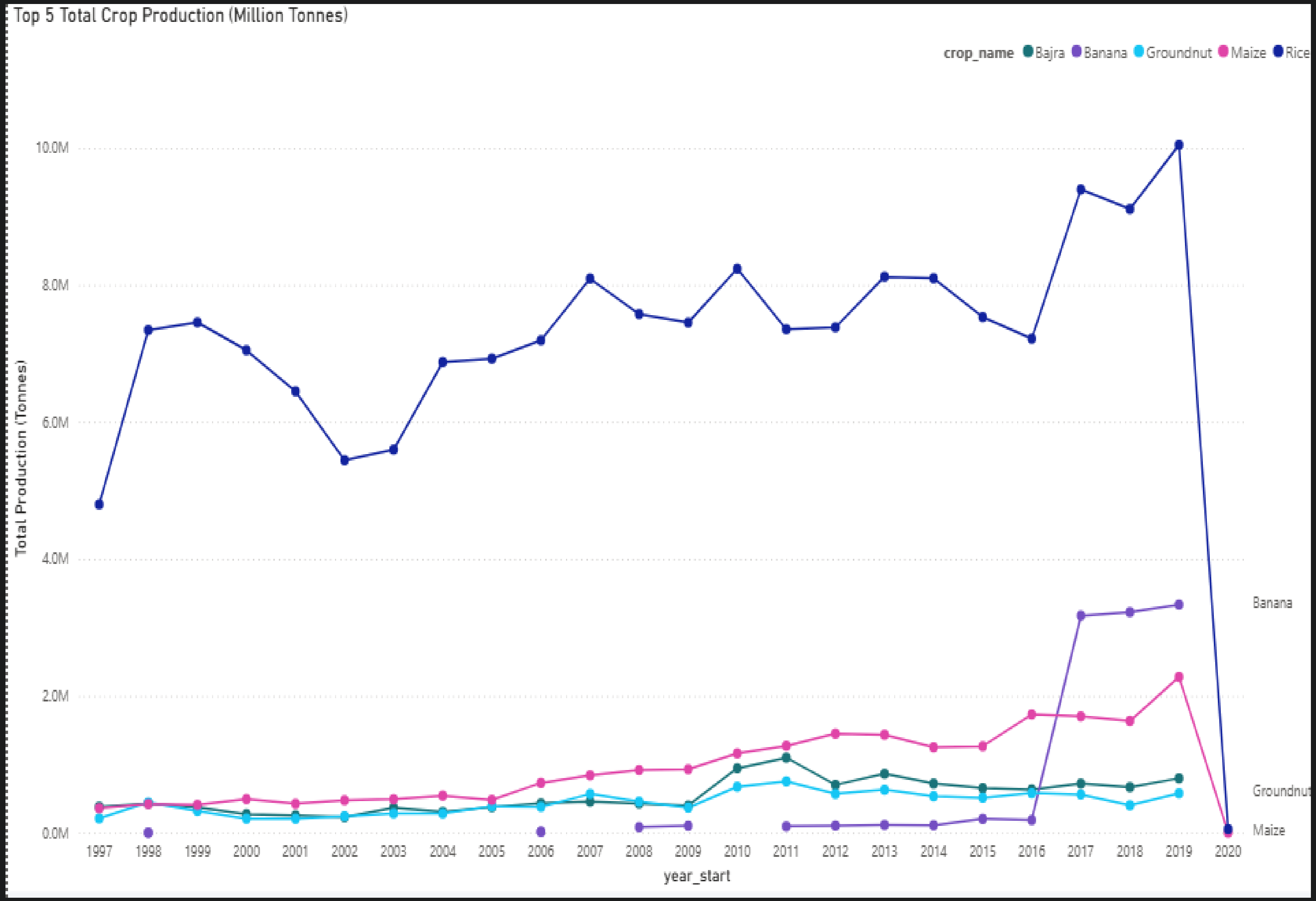
Enhanced Features

- **State-Level Drill-Through:** Navigate from overview to detailed state analysis
- **Dynamic KPI Cards:** Context-aware metrics for State, Season, and Year
- **Crop Production Breakdown:** Granular analysis by crop type
- **Multi-Year Trends:** Temporal patterns and growth trajectories

Technical Highlights

- DAX with `SELECTEDVALUE` for context-sensitive calculations
- `REMOVEFILTERS` to control filter context propagation
- Year-independent trend visualizations
- Cross-filtering across multiple dimensions

Analysis & Observation Achieved



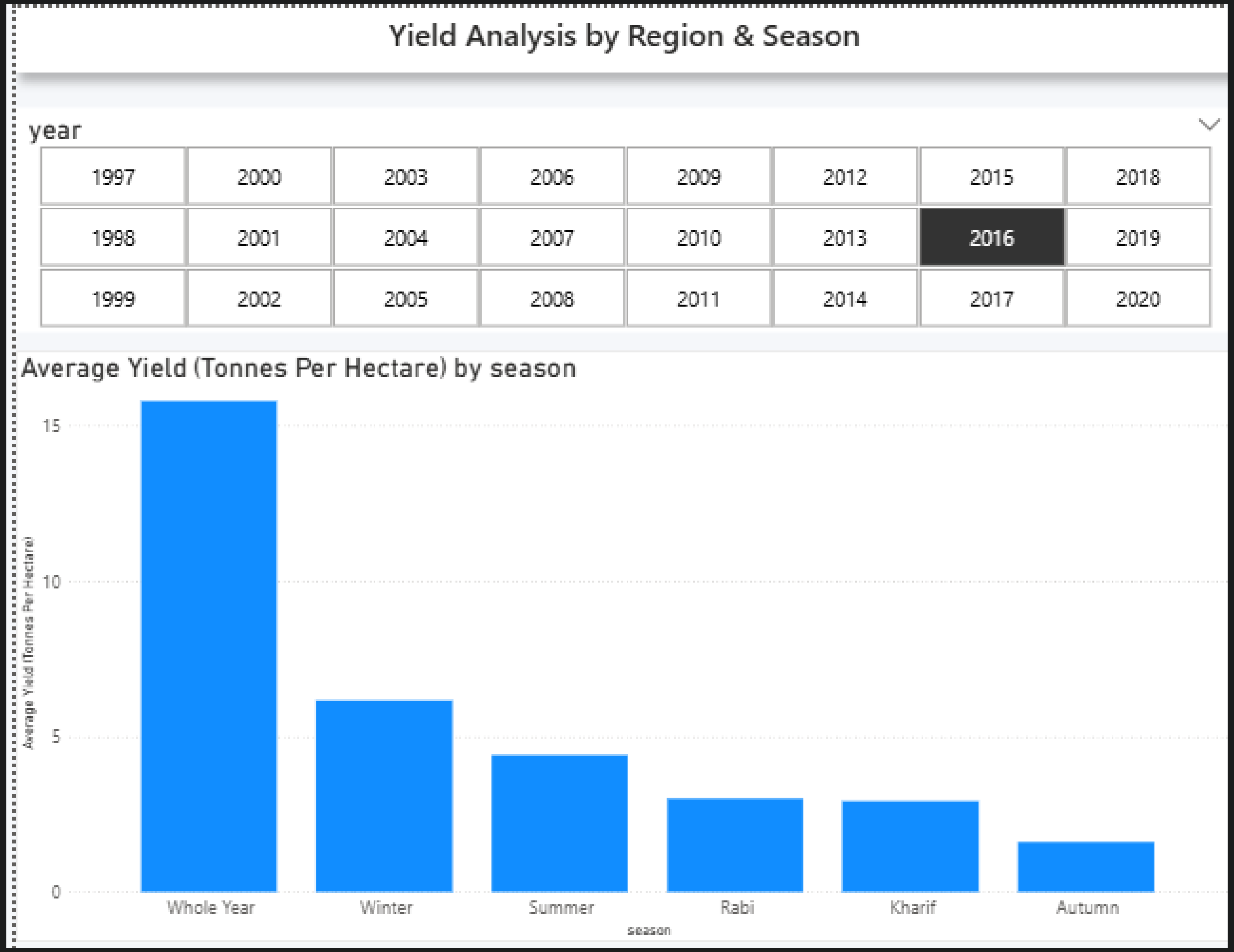
Business Insight

Identifies:

- Crops with stable growth
- Crops with volatile production
- Long-term production shifts

Helps policymakers understand **which crops are becoming dominant or declining**

Analysis & Observation Achieved

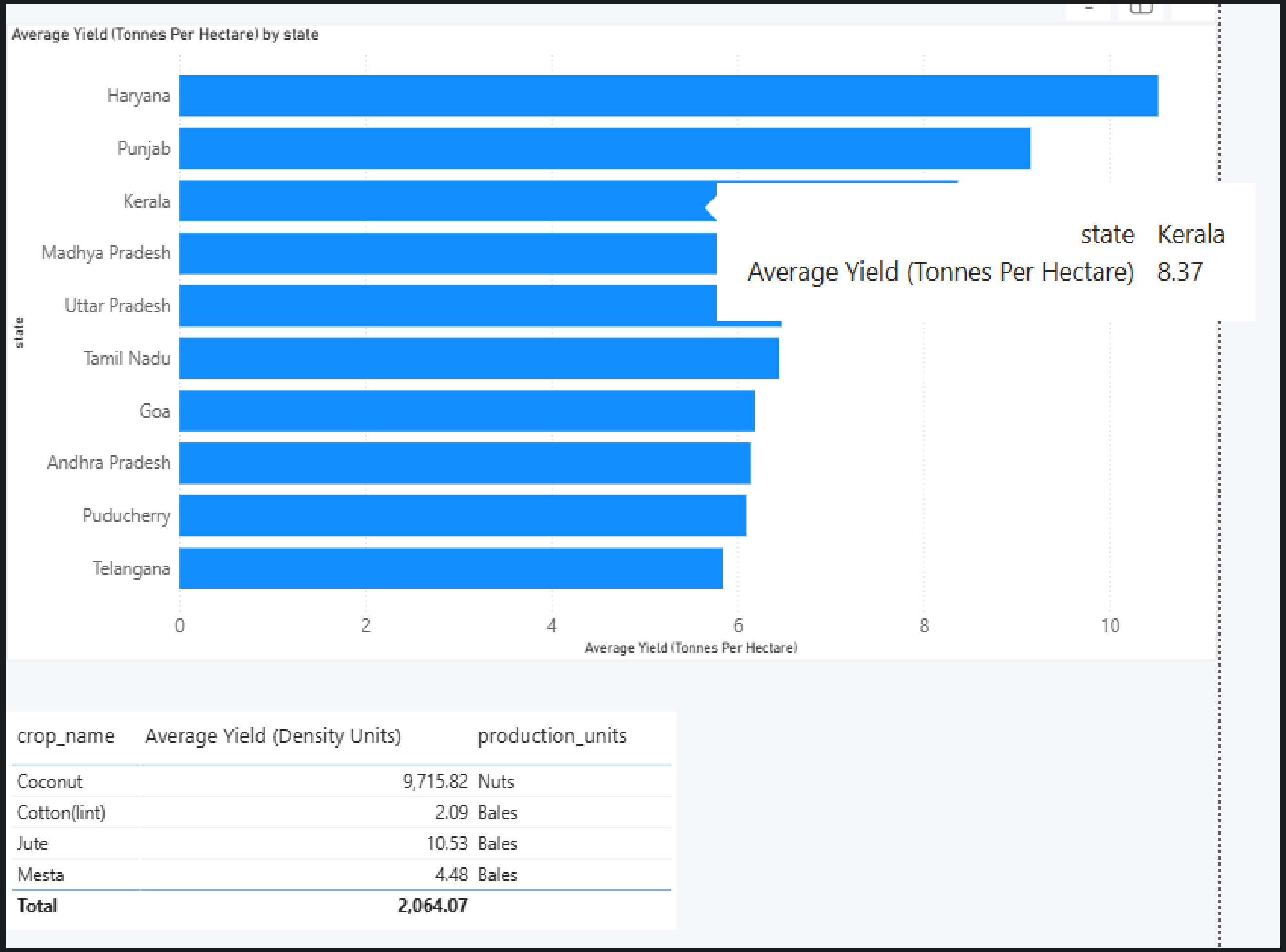


Yield Analysis by Region & Season

year

1997	2000	2003	2006	2009	2012	2015	2018
1998	2001	2004	2007	2010	2013	2016	2019
1999	2002	2005	2008	2011	2014	2017	2020

Average Yield (Tonnes Per Hectare) by season



Average Yield (Tonnes Per Hectare) by state

crop_name	Average Yield (Density Units)	production_units
Coconut	9,715.82	Nuts
Cotton(lint)	2.09	Bales
Jute	10.53	Bales
Mesta	4.48	Bales
Total	**2,064.07**	

Business Insight

Reveals:

Which seasons are most productive

Which states have higher land productivity

Key Technical Learnings & Outcomes

Engineering Depth

- Spark vs Pandas trade-offs for different data volumes
- Handling inconsistent units across datasets
- Complex yield recalculation logic

BI Optimization

- Managing filter context in Power BI
- DAX performance tuning
- Star schema modeling benefits

Pipeline Design

- Modular ETL architecture with Airflow
- Separation of concerns across layers
- Scalable, maintainable design patterns

Capstone Success

Delivered a fully automated, end-to-end analytics pipeline demonstrating enterprise-grade data engineering and BI implementation. All requirements successfully fulfilled with production-ready, scalable architecture.

> "This project demonstrates real-world data engineering and analytics practices at scale."