

Proyecto Big Data Fundatec

Marco Ferraro



Agenda

01

Contexto y Objetivo

02

Diseño

03

Integración de datos

04

Fase de predicción

05

Resultados

06

Preguntas

01

Contexto y Objetivo



Contexto y Objetivo

Open Data
Nepal

Home Datasets Organizations Suggest Data

Search

/ Datasets / District Wise Climate Data ...

District Wise Climate Data for Nepal

Followers
0

Social

Twitter

Facebook

License

Other (Public Domain) [Creative Commons](#)

Dataset

Groups

Activity Stream

District Wise Climate Data for Nepal

The dataset contains data on Nepal's climate on different parameters. These data were obtained from the NASA Langley Research Center (LaRC) POWER Project funded through the NASA Earth Science/Applied Science Program and extracted using NASA's power access API.

- Date (yyyy-mm-dd): 1981-01-01 to 2019-12-31
- Climate zone: NA (reference Briggs et al: <http://www.energycodes.gov>)
- Value for missing model data cannot be computed or out of model availability range: -999
- No of Districts: 62
- Community : SSE - Renewable Energy
- File Format: CSV

Parameters

1. PRECTOT: MERRA2 1/2x1/2 Precipitation (mm day-1)
2. PS: MERRA2 1/2x1/2 Surface Pressure (kPa)
3. QV2M: MERRA2 1/2x1/2 Specific Humidity at 2 Meters (g/kg)
4. RH2M: MERRA2 1/2x1/2 Relative Humidity at 2 Meters (%)
5. T2M: MERRA2 1/2x1/2 Temperature at 2 Meters (C)
6. T2MWET: MERRA2 1/2x1/2 Wet Bulb Temperature at 2 Meters (C)
7. T2M_MAX: MERRA2 1/2x1/2 Maximum Temperature at 2 Meters (C)
8. T2M_MIN: MERRA2 1/2x1/2 Minimum Temperature at 2 Meters (C)
9. T2M_RANGE: MERRA2 1/2x1/2 Temperature Range at 2 Meters (C)
10. TS: MERRA2 1/2x1/2 Earth Skin Temperature (C)
11. WS10M: MERRA2 1/2x1/2 Wind Speed at 10 Meters (m/s)
12. WS10M_MAX: MERRA2 1/2x1/2 Maximum Wind Speed at 10 Meters (m/s)
13. WS10M_MIN: MERRA2 1/2x1/2 Minimum Wind Speed at 10 Meters (m/s)
14. WS10M_RANGE: MERRA2 1/2x1/2 Wind Speed Range at 10 Meters (m/s)
15. WS50M: MERRA2 1/2x1/2 Wind Speed at 50 Meters (m/s)
16. WS50M_MAX: MERRA2 1/2x1/2 Maximum Wind Speed at 50 Meters (m/s)
17. WS50M_MIN: MERRA2 1/2x1/2 Minimum Wind Speed at 50 Meters (m/s)
18. WS50M_RANGE: MERRA2 1/2x1/2 Wind Speed Range at 50 Meters (m/s)



Información de Clima

El conjunto de datos contiene información sobre el clima de Nepal en diferentes parámetros. Estos datos fueron obtenidos del Centro de Investigación Langley de la NASA (LaRC).

Contexto y Objetivo



Información de Vegetales y Frutas

Este conjunto de datos contiene información oficial de precios para principales verduras y frutas en Nepal desde 2013 hasta 2021. El conjunto de datos incluye datos diarios de precios para cada verdura y fruta, así como los precios máximos, mínimos y promedio durante ese período.

Time Series Price Vegetables and Fruits

38 New Notebook Download (1 MB)

Data Card Code (7) Discussion (0)

kalimati_tarkari_dataset.csv (9.65 MB)

Detail Compact Column 7 of 7 columns

# SN	Commodity	Date	Unit	Minimum	Maximum	Average
0	Tomato Big(Nepali)	2013-06-16	Kg	35.0	40.0	37.5
1	Tomato Small(Local)	2013-06-16	Kg	26.0	32.0	29.0
2	Potato Red	2013-06-16	Kg	20.0	21.0	20.5
3	Potato White (Indian)	2013-06-16	Kg	15.0	16.0	15.5
4	Onion Dry	2013-06-16	Kg	28.0	30.0	29.0
5	Carrot(Local)	2013-06-16	Kg	30.0	35.0	32.5
6	Cabbage(Local)	2013-06-16	Kg	6.0	10.0	8.0
7	Cauli Local	2013-06-16	Kg	30.0	35.0	32.5
8	Raddish Red	2013-06-16	Kg	35.0	40.0	37.5
9	Raddish White(Local)	2013-06-16	Kg	25.0	30.0	27.5
10	Brinjal Long	2013-06-16	Kg	16.0	18.0	17.0


Data Explorer
Version 1 (9.65 MB)
kalimati_tarkari_dataset.csv

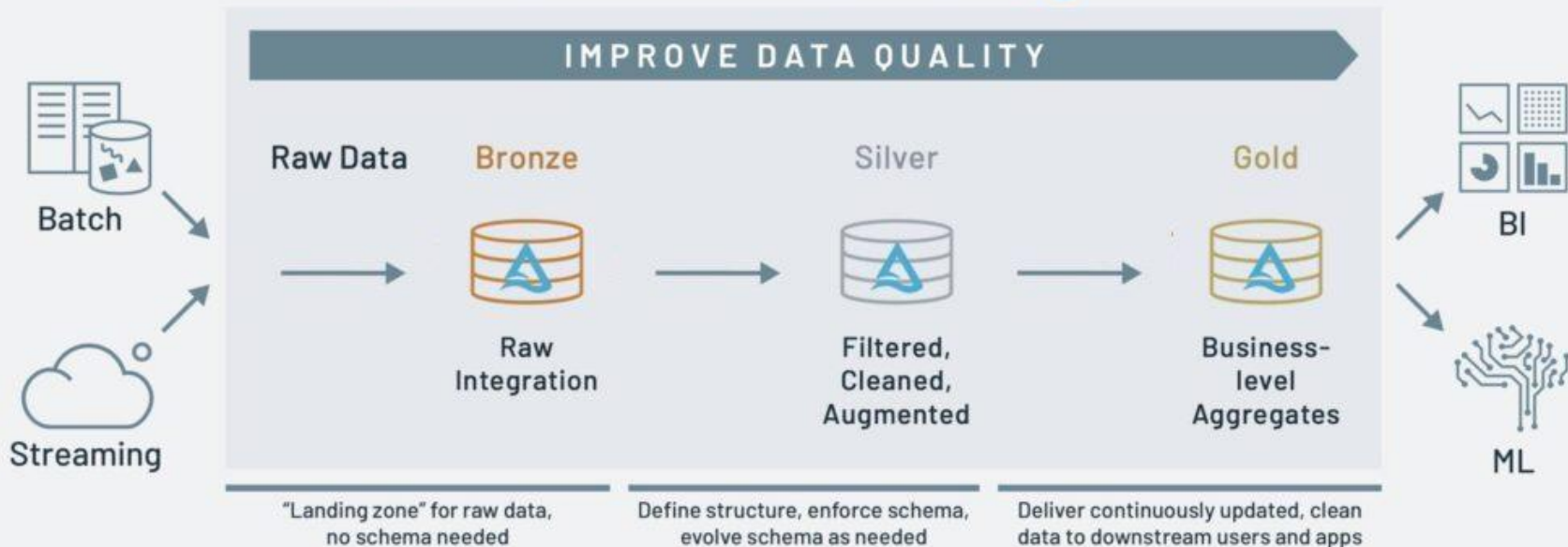
02

Diseño

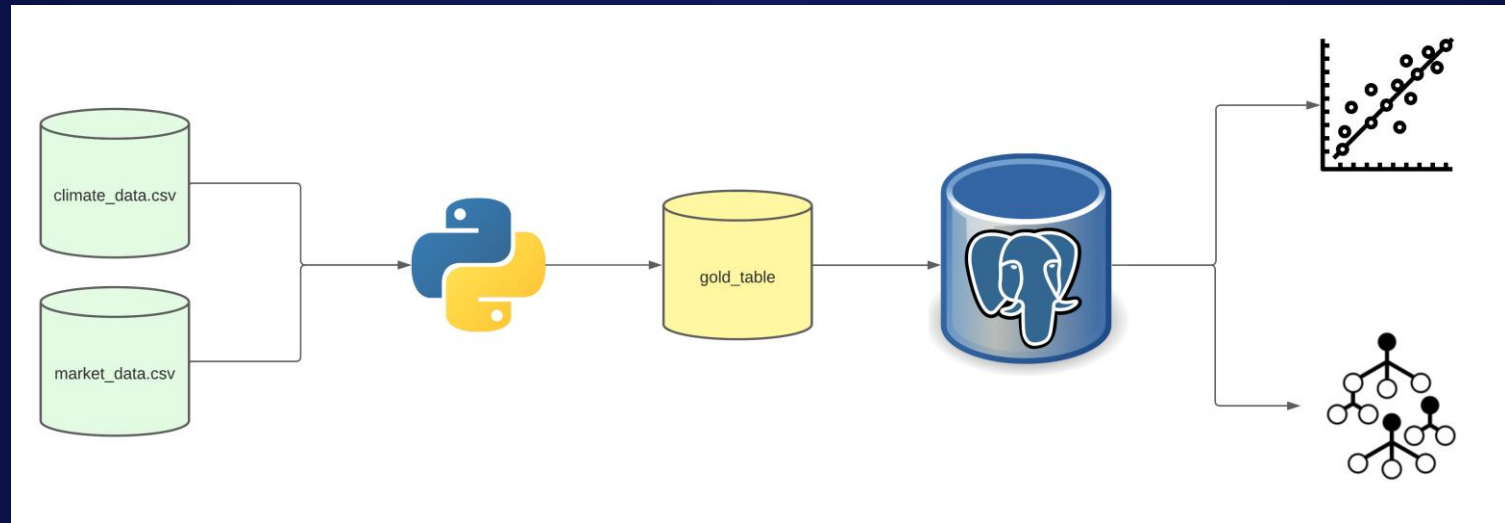


Diseño

Building reliable, performant data pipelines with  **DELTA LAKE**



Diseño



03

Integración de datos



Integración de datos

1. Cargar datos y borrar filas errneas

Primero se valida que no hayan filas duplicadas y se elimina filas duplicadas o filas que tengan null values.

2. Borrar Columnas innecesarias

Se eliminan columnas innecesarias que pueden realizar algún grado de ruido como "DISTRICT", "LAT", "LON", "PRECTOT".

3. Transformar datos de fechas

```
1 def transform_date_format(df, date_column, desired_format='yyyy-MM-dd'):  
2     df = df.withColumn(date_column, to_date(df[date_column], 'M/d/yyyy'))  
3  
4     df = df.withColumn(date_column, date_format(  
5         df[date_column], desired_format))  
6  
7     return df
```

4. Agregación de datos

```
1 def aggregate_dataframe(df, groupby_columns, avg_columns):  
2     avg_exprs = [F.avg(col).alias(f"AVG_{col}") for col in avg_columns]  
3     aggregated_df = df.groupBy(*groupby_columns).agg(*avg_exprs)  
4  
5     return aggregated_df
```

04

Fase de
predicción



Fase de predicción

1. Utilizamos un modulo implementado en Jupyter Notebook
2. Ingestamos datos sobre un servidor de PostgreSQL
3. Preposetamiento de datos: Hashing de Features y Escalado de datos
4. Usamos PySpark para medir el rendimiento de una Regresión Lineal y un algoritmo de Random Forrest.



05

Resultados



Resultados Linear Regression

```
Root Mean Squared Error (RMSE) on test data: 8.245760434257217  
Mean Squared Error (MSE) on test data: 67.99256513916177  
Mean Absolute Error (MAE) on test data: 2.5362482466187077  
Mean Squared Error (MSE) on test data: 67.99256513916177  
R-Squared ( $R^2$ ) on test data: 0.9878486384510242
```

Resultados Random Forrest

```
Root Mean Squared Error (RMSE) on test data for Random Forest: 41.262625106216795  
Mean Squared Error (MSE) on test data for Random Forest: 1702.6042306561928  
Mean Absolute Error (MAE) on test data for Random Forest: 34.22545413763678  
R-Squared ( $R^2$ ) on test data for Random Forest: 0.6957173252814526
```

Conclusiones

1. El modelo de regresión lineal es más interpretable, ya que podemos identificar el impacto individual de cada predictor en la variable de respuesta.
2. El Random Forest es menos interpretable en comparación, ya que se basa en múltiples árboles de decisión y no es fácil identificar cómo cada predictor afecta la predicción.
3. Docker provee un ambiente seguro para realizar diferentes módulos para ingesta de datos, procesamiento por batches, y analizar grandes volúmenes de datos.

06

Preguntas

