ENTERPRISE SCALE DATA BLENDING

Juniper Networks
Shalini Subramanian
Nov 9, 2018
WHO WE ARE

• Founded in 1996 in Sunnyvale, CA

• Leader in data center networking

ENTERPRISE  CLOUD  SERVICE PROVIDER
DATA LAKE ARCHITECTURE

KNIME + Big Data (Data Lake)

Ref: A Forest of Tools and Islands of Information – A Data Science Journey, Srinivas Attili, Director – Marketing Analytics & Data Science at Juniper Networks
LEAD UTILISATION

CHALLENGE:

• Lead Utilisation Report to be updated every week
• Excel Macros takes 8 hrs
• Manually triggering the update
• 300K rows dataset with about 60 columns
• Output report adds 40 extra columns
DATE/TIME TO STRING & VICE-VERSA

[Diagram showing a flow of operations for converting dates and times to strings and vice versa.]

© 2018 Juniper Networks
Comparing the value in Row N with Row (N-1)
```python
import pandas as pd
import csv
from pywebhdfs.webhdfs import PyWebHdfsClient
from datetime import datetime

# Copy input to output
output_table = input_table.copy()

output_table['Description'] = input_table['Description'].apply(lambda x: x.replace('
\n', '\t').replace('
', '\t').replace(' ', '\t'))

#input_table['Closed'] = input_table['Closed'].apply(lambda x: x if int(x) else 0)

output_table['Ingested_date'] = pd.datetime.now().strftime('%Y/%m/%d %H:%M:%S')
```

Python Script (1=1) (deprecated)
IN JOINS
LEAD UTILISATION

Outcome:

• Lead Utilisation Report to be updated every week
daily

• Excel Macros takes 8 hrs - 1 hour

• Manually triggering the update Scheduled
USE CASES

- Substitute for Excel macros
- Automation and scheduling
- Workflows showing business logic clearly
  - Batch Process
    - Customer Lifetime Value analysis
    - Modification of data list and then upload to datalake
  - Instream process
    - Data ingestion from S3/ Excel files into the datalake
    - Account hierarchies
    - Engagement score calculation
TYPICAL CONFIG SETTINGS

• Timeout for database operations (3000 in seconds)
• knime.ini file
  - knime.database.timeout=1000
  - Xmx4096m (Increase the java heap space memory for local operations)
CONTINUING TO USE KNIME

• KNIME integrations with various tools (ranging from ETL to Viz)
• Incredibly useful when processing a file for checks/validations/correctness
• Able to set up a data pipe/process in an automated manner.