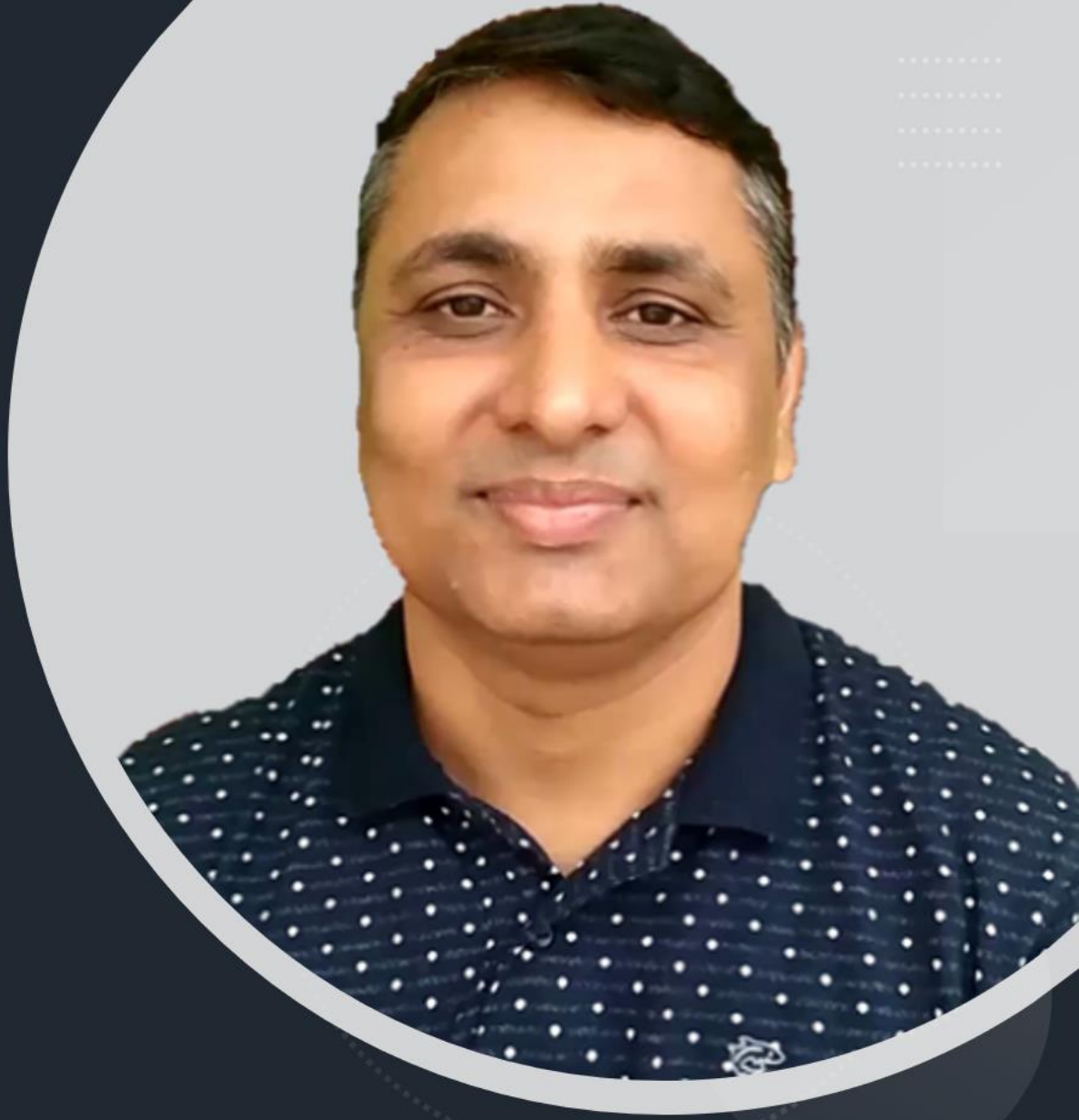
ScholarNest

Spark Azure Databricks

Databricks Spark Certification and beyond

Instructor: Prashant Kumar Pandey



Absolute Beginner to Specialization in Apache Spark and Azure Databricks

Assignments

Assignment - 1

Design Hadoop Cluster Architecture to offer the following total capacity for your Spark application team.

- CPU Capacity - 240 CPU
- Memory Capacity - 2048 GB
- Storage Capacity - 400 TB

You can make appropriate assumptions for master and worker node capacity.

Assignment - 2

- You have the following two teams.
    1. Data Engineering Team
    2. Data Analysis Team

The Data Engineering team collects and prepares the data for the Data Analysis team.

The Data Analysis team primarily uses SQL, Reporting, and BI Tools to analyze data and create dashboards.

Design a platform architecture using Databricks Cloud so both teams can work smoothly.

Assignment - 3

Write a spark-submit command to run an application with the following files.

1. ctr_analysis_main.py
2. analyse_libs.zip
3. app_configs.properties

You also want to set the following configurations

1. Application name should be CTR-ANALYSIS
2. spark.sql.autoBroadcastJoinThreshold should be set to 250 MB

Assignment - 4

Create a Spark application on your local machine using IDE to do the following

1. Create a Spark session and set your application name.
2. Read a CSV data file to create a data frame.
3. The Data file location should be given as a command-line argument.
4. Show the data frame

Assignment - 5

You are asked to create the below data frame using five method listed below.

id	name	dob
101	Prashant	25-10-1977
102	Tanisk	28-01-1982
103	Jahid	14-09-1986

The column names of the data frame must match as given in the example. The data types of the columns must match as defined below.

- id -> String
- name -> String
- dob -> Date

1. Create a data frame using toDF() method without an schema definition.
2. Create a data frame using a schema that only defines the column names and no data types.
3. Create a data frame using a schema that defines the column name and data types.
4. Create a list of Row() and a schema. Use the list of Row() and schema to define a data frame.
5. Create a Spark RDD and use it to define a data frame.

Assignment - 6

Load the flight_time.json data and create a data frame as explained below.

1. Define the schemes to enforce the column names and data types
2. Print schema and validate if you loaded it correctly
3. Correct the schema of some columns, such as the CANCELLED column, if required
4. Save the data frame as a Spark Table
5. Describe the table to check the data types are correctly defined

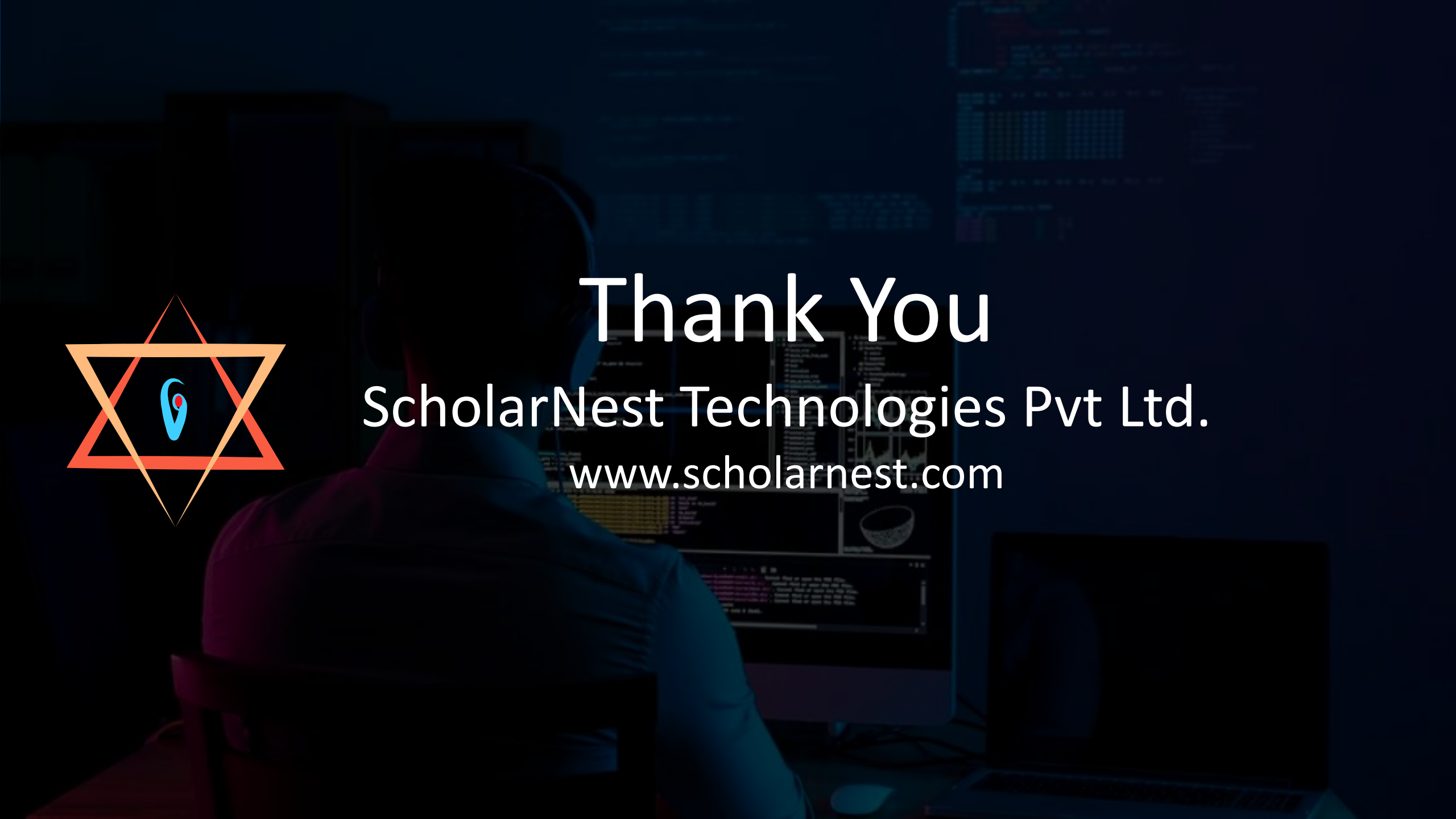
Assignment - 7

You are given a data file at the following location.

Data file: /databricks-datasets/learning-spark-v2/sf-fire/sf-fire-calls.csv

1. Define a schema DDL for this data file
2. Load the data using schema to create a data frame
3. The Column names and data types must match the given schema.
4. Correct the data frame to fix date column data types.
5. Save the corrected data frame as a Spark table
6. Create an external table on the above defined table

```
root
 |-- CallNumber: integer
 |-- UnitID: string
 |-- IncidentNumber: integer
 |-- CallType: string
 |-- CallDate: string
 |-- WatchDate: string
 |-- CallFinalDisposition:
 |-- AvailableDtTm: string
 |-- Address: string
 |-- City: string
 |-- Zipcode: integer
 |-- Battalion: string
 |-- StationArea: string
 |-- Box: string
 |-- OrigPriority: string
 |-- Priority: string
 |-- FinalPriority: integer
 |-- ALSUnit: boolean
 |-- CallTypeGroup: string
 |-- NumAlarms: integer
 |-- UnitType: string
 |-- UnitSequenceInCallDispatch: integer
 |-- FirePreventionDistrict: string
 |-- SupervisorDistrict: string
 |-- Neighborhood: string
 |-- Location: string
 |-- RowID: string
 |-- Delay: double
```



Thank You

ScholarNest Technologies Pvt Ltd.

www.scholarnest.com