Provenance Filter by DOID Numeric

Purpose

This script filters provenance data from a Parquet file (`study_refs.parquet`) based on a numeric DOID value provided by the user. It outputs the filtered dataset as a new Parquet file for downstream analysis.

Key Features

- Dynamically filters data based on a specific DOID numeric value.
- Efficiently processes large datasets using Apache Spark.
- Saves filtered results in Parquet format for optimized storage and querying.

Requirements

- **Apache Spark**: Ensure PySpark is installed and properly configured.
- **Input File**:
  - Default: `study_refs.parquet`
  - Format: Parquet
  - Required Column: `doid_uniprot`
- **Command-Line Argument**:
  - `DOID_number`: The numeric part of the DOID for filtering.

Workflow

Step 1: Initialize Spark Session

The script initializes a Spark session with memory configurations for handling large datasets:

- **Driver Memory**: 8 GB
- **Executor Memory**: 16 GB

```
spark = SparkSession.builder \
    .appName("Filter DOID Provenance") \
    .config("spark.driver.memory", "8g") \
    .config("spark.executor.memory", "16g") \
    .getOrCreate()
```

Step 2: Read Input File

The dataset is loaded from the input Parquet file specified in the script or dynamically provided. The file must contain the `doid_uniprot` column.

```
df = spark.read.parquet(input_file)
```

Step 3: Construct DOID Search Value

The numeric DOID provided as a command-line argument is used to construct a search string in the format `DOID:<DOID_number>`.

```
doid_search_value = f"DOID:{doid_number}"
```

Step 4: Filter Data

The script filters the rows where the `doid_uniprot` column contains the constructed DOID value.

```
filtered_df = df.filter(col("doid_uniprot").contains(doid_search_value))
```

Step 5: Save Filtered Results

The filtered data is saved as a new Parquet file named `<DOID_number>_provenance.parquet` in overwrite mode.

```
filtered_df.write.mode("overwrite").parquet(output_file)
```

Step 6: Stop Spark Session

The Spark session is stopped to release resources after the operation completes.

```
spark.stop()
```

Execution

1. Place the input file (`new_transformed_study_refs.parquet`) in the working directory or specify its path in the script.
2. Run the script with the desired DOID number as a command-line argument:

```
python provenance_filter.py <DOID_number>
```

3. Example:

```
python provenance_filter.py 9352
```

Output

- **File Format**: Parquet
- **File Name**: `<DOID_number>_provenance.parquet`
- **Content**: Rows from the input dataset where `doid_uniprot` contains the specified DOID numeric value.

Performance Optimizations

- **Selective Filtering**: Uses `contains` function to dynamically search within the `doid_uniprot` column.
- **Memory Configuration**: Configured driver and executor memory for efficient processing of large datasets.

Future Improvements

- Add validation for the existence of the input file and column `doid_uniprot`.
- Allow dynamic input file paths through command-line arguments or a configuration file.
- Extend filtering capabilities to include additional criteria or multiple DOID values.

This documentation ensures clarity and usability, enabling users to understand, execute, and extend the script as needed.