

Experiment 9

Aim: Graph Analytics with GraphX: Explore graph-structured data using Apache Spark's GraphX library, and perform graph analytics tasks such as computing centrality measures or detecting communities.

Spark GraphX Analytics:

```
C:\Users\rawat\Documents\7 SEMESTER\Big Data Analytics\Lab\Experiment 9\Codes>spark-submit --packages graphframes:graphframes:0.8.2-spark3.1-s_2.12 Spark_GraphX_Centrality_Measure.py
:: loading settings :: url = jar:file:/C:/Spark/jars/ivy-2.5.1.jar!/org/apache/ivy/core/settings/ivysettings.xml
Ivy Default Cache set to: C:\Users\rawat\ivy2\cache
The jars for the packages stored in: C:\Users\rawat\ivy2\jars
graphframes#graphframes added as a dependency
:: resolving dependencies :: org.apache.spark#spark-submit-parent-e4b3ee5d-0aff-42de-92f6-422ea4a9d24b;1.0
  confs: [default]
    found graphframes#graphframes;0.8.2-spark3.1-s_2.12 in spark-packages
    found org.slf4j#slf4j-api;1.7.16 in central
downloading https://repos.spark-packages.org/graphframes/graphframes/0.8.2-spark3.1-s_2.12/graphframes-0.8.2-spark3.1-s_2.12.jar ...
[SUCCESSFUL ] graphframes#graphframes;0.8.2-spark3.1-s_2.12|graphframes.jar (5358ms)
downloading https://repo1.maven.org/maven2/org/slf4j/slf4j-api/1.7.16/slf4j-api-1.7.16.jar ...
[SUCCESSFUL ] org.slf4j#slf4j-api;1.7.16|slf4j-api.jar (1711ms)
:: resolution report :: resolve 8439ms :: artifacts dl 7106ms
  :: modules in use:
    graphframes#graphframes;0.8.2-spark3.1-s_2.12 from spark-packages in [default]
    org.slf4j#slf4j-api;1.7.16 from central in [default]
-----
| conf | number | search | dwnlded | evicted | artifacts |
|-----|-----|-----|-----|-----|-----|
| default | 2 | 2 | 2 | 0 | 2 | 2 |
-----
:: retrieving :: org.apache.spark#spark-submit-parent-e4b3ee5d-0aff-42de-92f6-422ea4a9d24b
  confs: [default]
  2 artifacts copied, 0 already retrieved (281kB/323ms)
24/11/24 20:12:29 INFO SparkContext: Running Spark version 3.5.2
24/11/24 20:12:29 INFO SparkContext: OS info Windows 11, 10.0, amd64
24/11/24 20:12:29 INFO SparkContext: Java version 1.8.0_421
24/11/24 20:12:29 INFO ResourceUtils: ==============================================================
```

Fig 1: Installing Required Packages

```
24/11/24 20:12:31 INFO BlockManagerMaster: Registered BlockManager Blo
24/11/24 20:12:31 INFO BlockManager: Initialized BlockManager: BlockMa
24/11/24 20:12:32 INFO SharedState: Setting hive.metastore.warehouse.d
24/11/24 20:12:32 INFO SharedState: Warehouse path is 'file:/C:/Users/
C:\Spark\python\lib\pyspark.zip\pyspark\sql\dataframe.py:168: UserWarn
sparkSession instead.
Vertices:
24/11/24 20:12:34 INFO CodeGenerator: Code generated in 200.2087 ms
24/11/24 20:12:34 INFO SparkContext: Starting job: showString at Nativ
24/11/24 20:12:35 INFO DAGScheduler: Got job 0 (showString at NativeMe
24/11/24 20:12:35 INFO DAGScheduler: Final stage: ResultStage 0 (showS
24/11/24 20:12:35 INFO DAGScheduler: Parents of final stage: List()
24/11/24 20:12:35 INFO DAGScheduler: Missing parents: List()
24/11/24 20:12:35 INFO DAGScheduler: Submitting ResultStage 0 (UserDefi
```

Fig 2: Dataset Schema

```
24/11/24 20:12:41 INFO TaskSched
24/11/24 20:12:41 INFO DAGSched
24/11/24 20:12:42 INFO CodeGene
+---+-----+
| id | name |
+---+-----+
| 1 | Alice |
| 2 | Bob |
| 3 | Charlie |
| 4 | David |
| 5 | Eve |
+---+-----+

Edges:
24/11/24 20:12:42 INFO SparkCon
```

Fig 3: First Rows and Classes of Dataset

```
24/11/24 20:12:47 INFO TaskSchedulerImpl: Kill
24/11/24 20:12:47 INFO DAGScheduler: Job 5 fi
+---+-----+
|src|dst|
+---+-----+
| 1 | 2 |
| 2 | 3 |
| 3 | 4 |
| 4 | 5 |
| 1 | 3 |
+---+-----+

C:\Spark\python\lib\pyspark.zip\pyspark\sql\d
Degree Centrality:
24/11/24 20:12:48 INFO CodeGenerator: Code ge
```

Fig 4: Source and Destination

```
24/11/24 20:12:55 INFO DAGScheduler: Job 7 f
24/11/24 20:12:55 INFO TaskSchedulerImpl: Ki
24/11/24 20:12:55 INFO DAGScheduler: Job 7 f
+---+-----+
| id | degree |
+---+-----+
| 1 | 2 |
| 2 | 2 |
| 3 | 3 |
| 4 | 2 |
| 5 | 1 |
+---+-----+

24/11/24 20:12:55 INFO ConnectedComponents$:
```

Fig 5: Source ID and Degree