Lënda: Big Data

NoSQL, Manipulim me Big Data dhe Analiza e Grafeve

Punuan:
Endrit Kastrati
Erza Gashi
Jon Kuçi
Kaltrina Krasniqi
Mirgeta Gashi

Prof. Dr. Vigan Raça
Msc. Ass. Rafet Duriqi

Përmbajtja

Hyrje

NoSQL

Manipulimi me Big Data - Spark

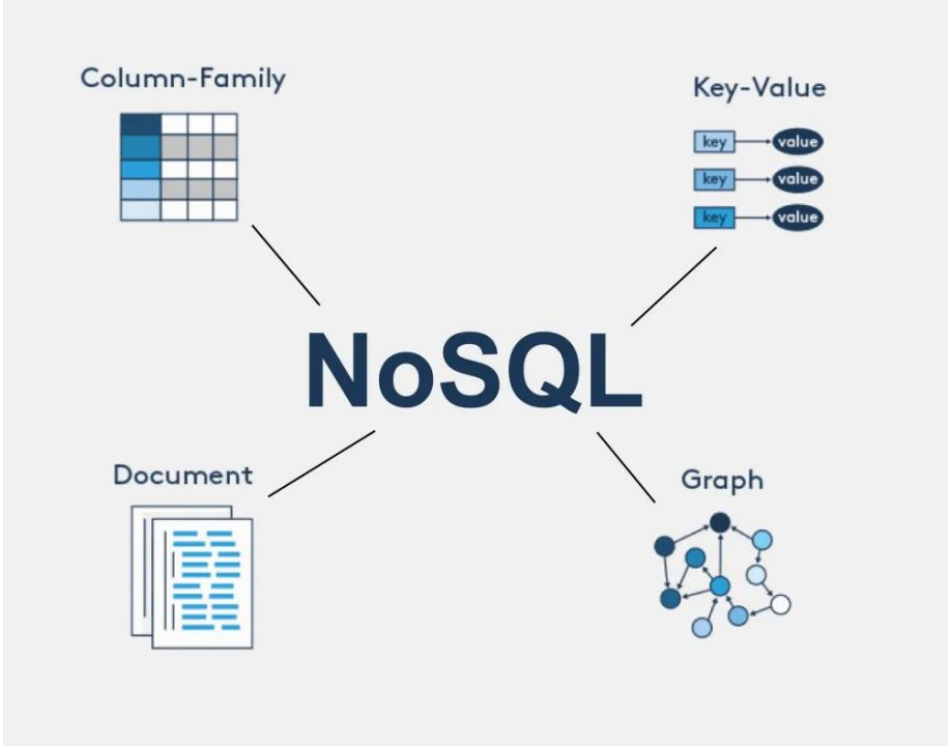
Analiza e grafeve

NoSQL

NoSQL qëndron për **"Not Only SQL"** – një alternativë ndaj databazave tradicionale relacionale..

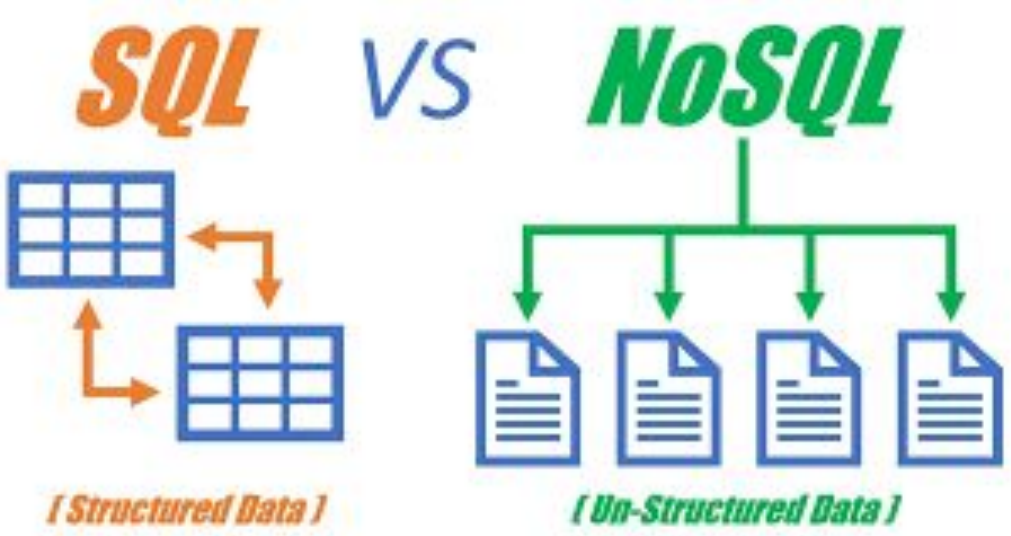
Llojet kryesore të NoSQL:

- **Dokumente** (MongoDB) – ruan të dhëna si JSON/BSON
- **Key-Value Stores** (Redis, DynamoDB)
- **Graph Database** (Gephi)
- **Column Stores**(Cassandra, HBase)



SQL vs NoSQL

Structured Query Language v Not only Structured Query Language



SQL:

- Të dhëna të strukturuara në tabela.
- Struktura fikse (me skemë).
- I përshtatshëm për joins komplekse dhe transaksione.
- Query/Komanda specifike SQL.

NoSQL:

- Modele fleksibile të të dhënave (JSON, çelës-vlerë, grafik, etj.)
- Pa skemë ose me skemë dinamike.
- I përshtatshëm për big data dhe aplikacione në kohë reale.

MongoDB

MongoDB është një **bazë e të dhënave NoSQL dokument-based**, e cila ruan informacionin në format **JSON/BSON.**

Avantazhet:

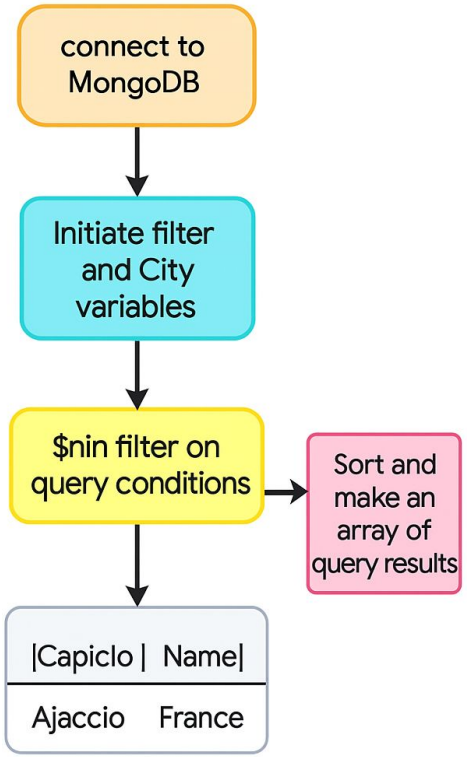
- **Fleksibilitet i lartë i skemës** – përshtatet me të dhëna të ndryshueshme.
- **Shkallëzim horizontal** – i përshtatshëm për sisteme të mëdha dhe të shpërndara.
- **Query të fuqishme** përmes operatorëve si `$in`, `$nin`, `$sort`, etj.
- I integrueshëm me shumë gjuhë programimi dhe teknologji moderne.

Disavantazhet:

- Nuk ka JOIN klasik (si në SQL)
- Kërkon logjikë të shtuar për ndërlidhje të të dhënave
- Më pak i përshtatshëm për transaksione komplekse

Manipulimi me Big Data - MongoDB - Mondial

Eksporti i mondial nga SQL në csv file është bërë me veglën: [Github Link]



Query 2: Te listohen te gjitha kryeqytetet ne te cilat nuk kalon asnje lum si dhe shtetet e tyre nuk kane dalje ne det

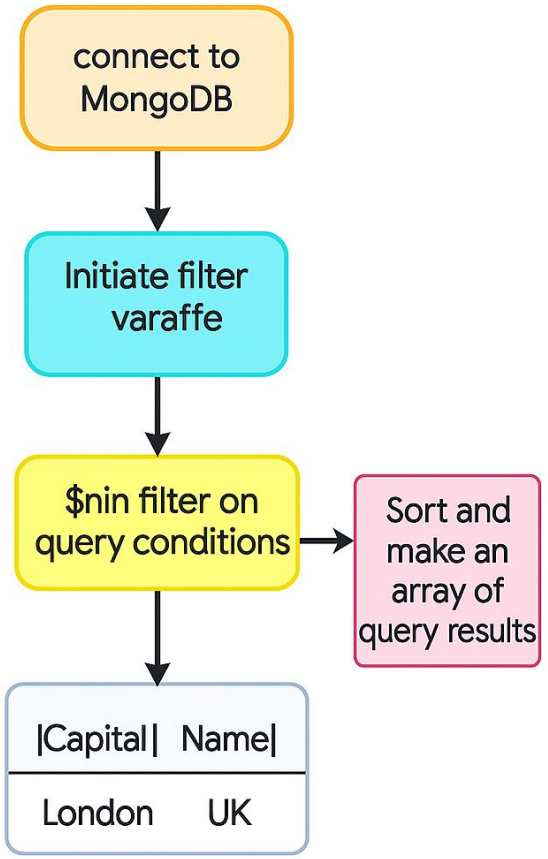
```
mondial> db.country.find( { code: { $nin: seaCountries }, capital: { $nin: riverCities } }, { _id: 0, capital: 1, name:
 1 } ).sort({ capital: 1 });
[
  { name: 'Ethiopia', capital: 'Addis Ababa' },
  { name: 'Andorra', capital: 'Andorra la Vella' },
  { name: 'Turkmenistan', capital: 'Ashgabat' },
  { name: 'Azerbaijan', capital: 'Baku' },
  { name: 'Kyrgyzstan', capital: 'Bishkek' },
  { name: 'Moldova', capital: 'Chişinău' },
  { name: 'Tajikistan', capital: 'Dushanbe' },
  { name: 'Botswana', capital: 'Gaborone' },
  { name: 'Zimbabwe', capital: 'Harare' },
  { name: 'Afghanistan', capital: 'Kabul' },
  { name: 'Uganda', capital: 'Kampala' },
  { name: 'Nepal', capital: 'Kathmandu' },
  { name: 'Rwanda', capital: 'Kigali' },
  { name: 'Bonaire', capital: 'Kralendijk' },
  { name: 'Bolivia', capital: 'La Paz' },
  { name: 'Malawi', capital: 'Lilongwe' },
  { name: 'Zambia', capital: 'Lusaka' },
  { name: 'Lesotho', capital: 'Maseru' },
  { name: 'Eswatini', capital: 'Mbabane' },
  { name: 'Belarus', capital: 'Minsk' }
]
Type "it" for more
mondial> it
[
  { name: 'Burkina Faso', capital: 'Ouagadougou' },
  { name: 'Kosovo', capital: 'Prishtine' },
  { name: 'West Bank', capital: 'Ramallah' },
  { name: 'San Marino', capital: 'San Marino' },
  { name: 'Bhutan', capital: 'Thimphu' },
  { name: 'Uzbekistan', capital: 'Toshkent' },
  { name: 'Vatican City', capital: 'Vatican City' }
]
```

Showing rows: 1 to 27 Page No: 1 of 1

	capital character varying (50)	countryname character varying (50)
1	Addis Ababa	Ethiopia
2	Andorra la Vella	Andorra
3	Ashgabat	Turkmenistan
4	Baku	Azerbaijan
5	Bishkek	Kyrgyzstan
6	Chişinău	Moldova
7	Dushanbe	Tajikistan
8	Gaborone	Botswana
9	Harare	Zimbabwe
10	Kabul	Afghanistan
11	Kampala	Uganda
12	Kathmandu	Nepal
13	Kigali	Rwanda
14	Kralendijk	Bonaire
15	La Paz	Bolivia
16	Lilongwe	Malawi
17	Lusaka	Zambia
18	Maseru	Lesotho

Manipulimi me Big Data - MongoDB - Mondial

Eksporti i mondial nga SQL në csv file është bërë me veglën: [Github Link]



Query/View 4: Te listohen kryeqytetet e shteteve te cilat nuk jane antare ne asnje organizate boterore

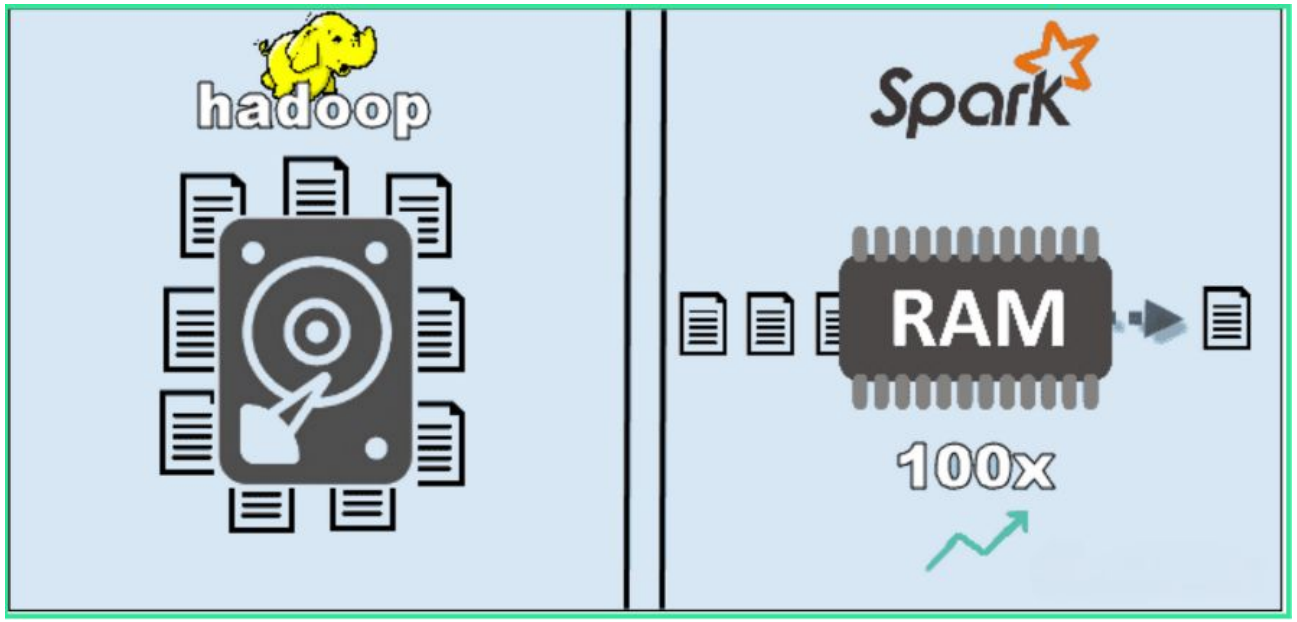
```
mondial> result
[
  { name: 'Bonaire', capital: 'Kralendijk' },
  { name: 'Ceuta', capital: 'Ceuta' },
  { name: 'Christmas Island', capital: 'Flying Fish Cove' },
  { name: 'Cocos Islands', capital: 'West Island' },
  { name: 'Gaza Strip', capital: '' },
  { name: 'Melilla', capital: 'Melilla' },
  { name: 'Svalbard', capital: 'Longyearbyen' },
  { name: 'West Bank', capital: 'Ramallah' }
]
```

Manipulimi me Big Data

Manipulimi i big data është bërë me spark

Spark

→ I fuqishëm për analiza të mëdha në memorie
→ E përshtatshme për punë analitike
→ Ka performancë shumë të lartë dhe integrohet lehtësisht me mjete si Pandas, Matplotlib dhe bibliotekat e Python-it.

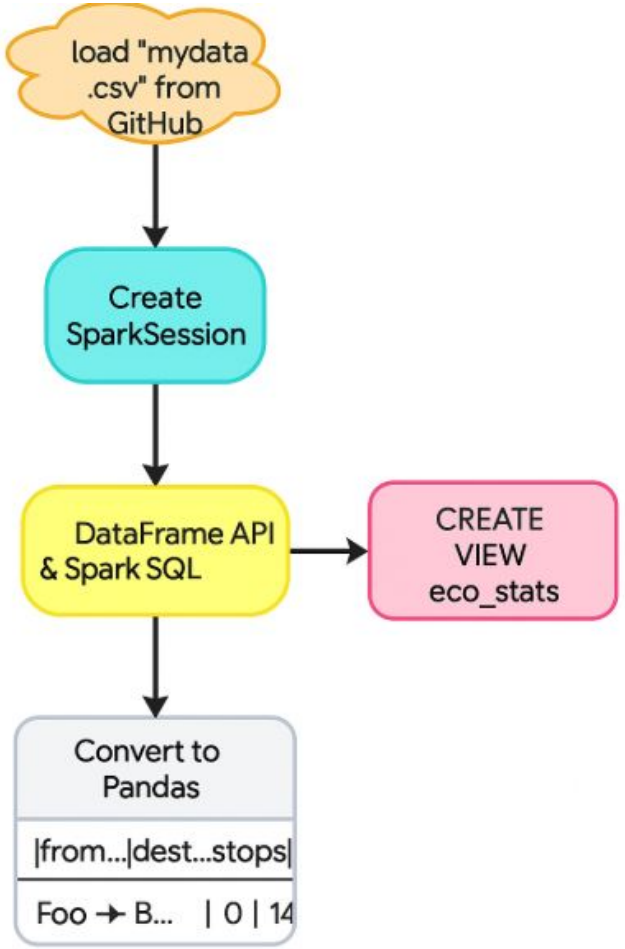


Hadoop

→ Platformë për përpunim batch, me bazë në MapReduce.
→ Kërkon konfigurim kompleks dhe infrastrukturë të shpërndarë (cluster ose HDFS lokal).
→ Më i përshtatshëm për procese të mëdha të njëpasnjëshme, jo për analiza interaktive apo zhvillim të shpejtë.

Manipulimi me Big Data - Spark - Mondial

Eksporti i mondial nga SQL në csv file është bërë me veglën: Github Link



Query 5: Te listohen te gjithe lumenjte te cilet kalojne neper vendet antare te NATO-s dhe EU-se perjashtuar Suedinte dhe Francen

Query 3: Te listohen te gjitha kryeqytetet e shteteve anetare te NATO-s ne te cilat kalon te pakten nje lum

Manipulimi me Big Data - Spark - Flight Dataset

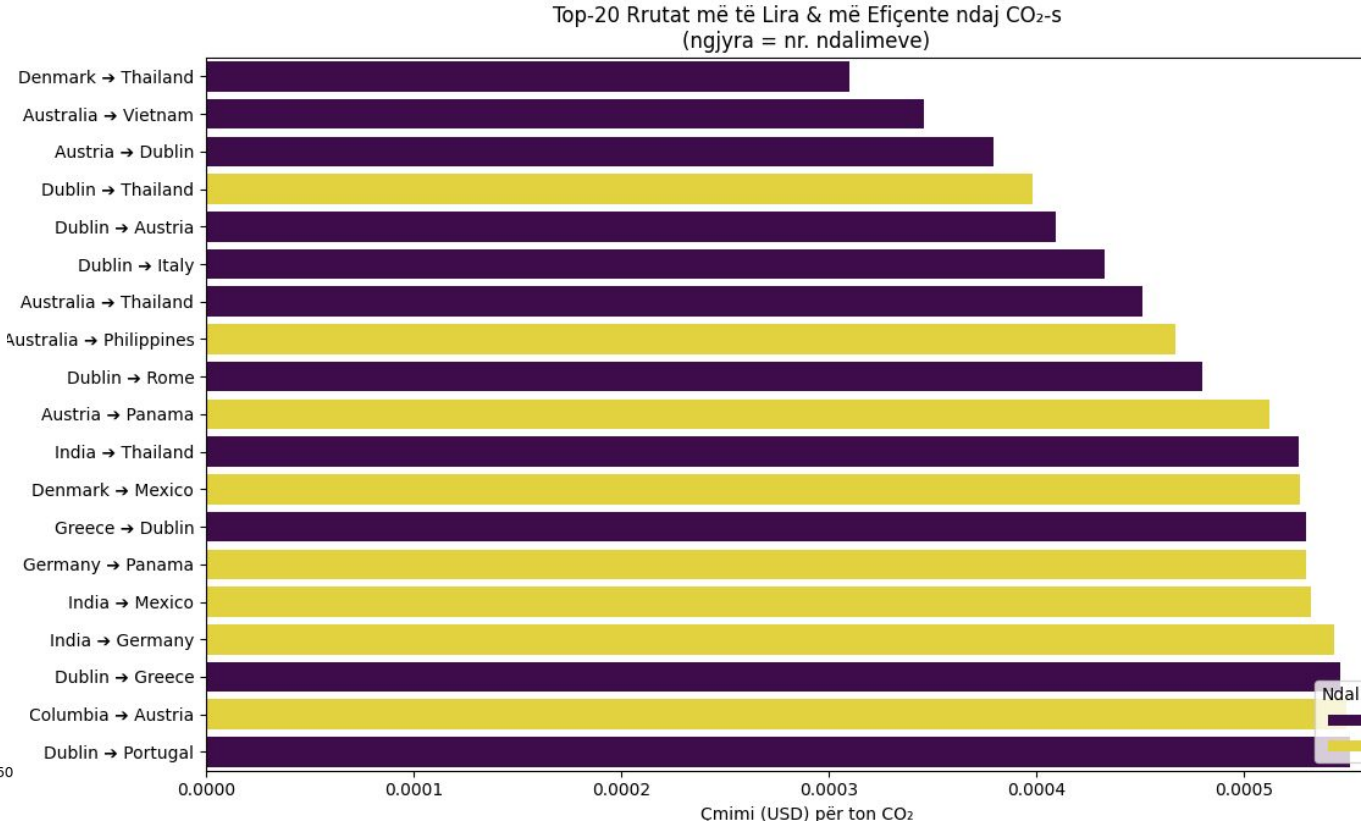
Për të testuar pjesën e “Big Data”, është përdorur një dataset real me të dhëna të fluturimeve ndërkombëtare, që përmban **~ 1 milion rekorde** mbi çmime, ndalesa dhe detaje teknike të linjave ajrore dhe size mbi **200 MB**. Flight Dataset

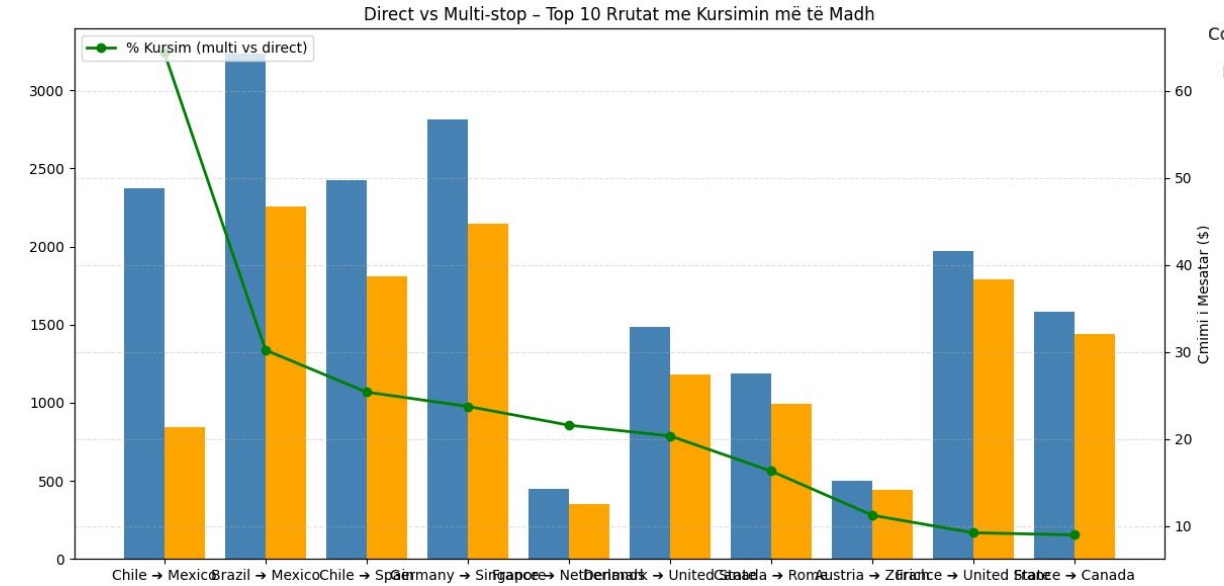
Ky dataset është analizuar me Spark për të nxjerrë statistika rreth efikasitetit të çmimit dhe numrit të ndalimeve në rrugë të ndryshme.

Manipulimi me Big Data - Spark - Flight Dataset

VIzualizimet me **mathplotlib**

Query 1: Fluturimet më të Lira kundrejt më Eco-Friendly sipas Rrugës dhe Numrit të Ndalimeve

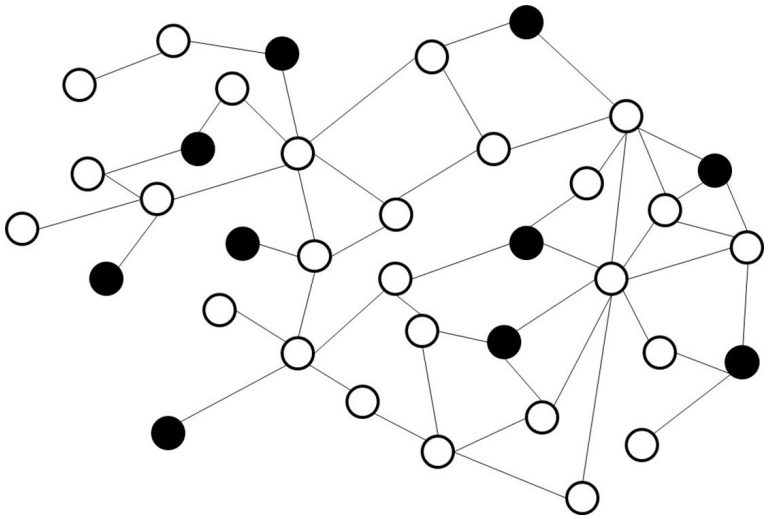




Query 2: Sa % MË LIRË apo MË SHUMË kushton të shtosh ndalesa?

Analiza e Grafeve - Social Network Analysis

- SNA - metodë për të analizuar struktura sociale për të gjetur marrëdhëniet/ndërveprimet ndërmjet entiteteve.
- Përdor grafe me nodes dhe edges për të modeluar rrjetin.
- Zbulon nyjet më të rëndësishme përmes metrikave si Degree, Closeness dhe Betweenness Centrality.



Sociogram

→ $Degree(v) = k_v$

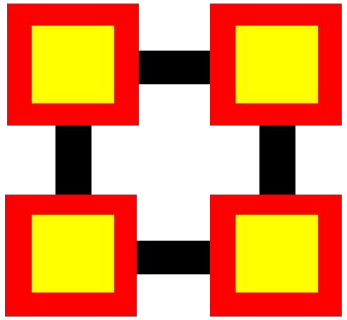
→ $Betweenness(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}}$

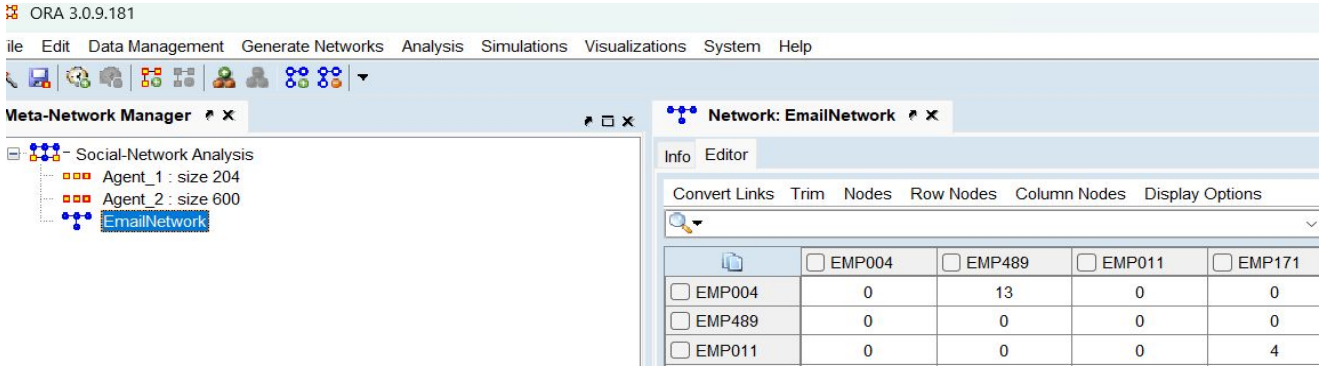
→ $Closeness(v) = \frac{n-1}{\sum_{u \neq v} d(u,v)}$

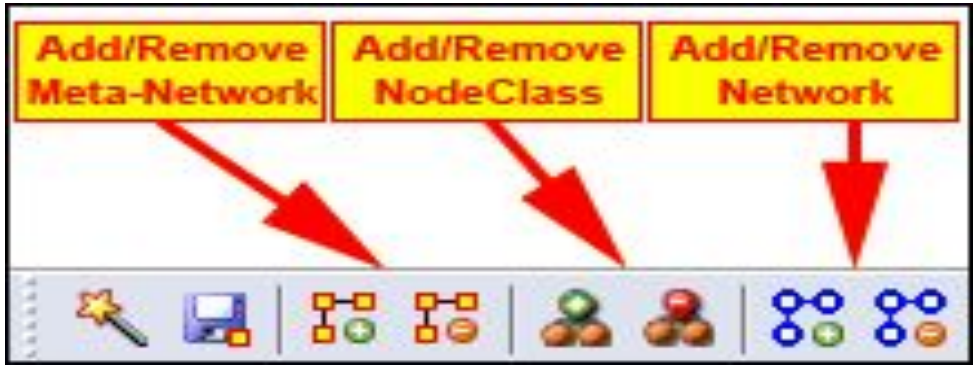
	C1	C2	C3	C4
C1	-	1	1	0
C2	1	-	0	1
C3	1	0	-	0
C4	0	1	0	-

Analiza e Grafeve - ORALite

- Mjet për Analizën dhe Vizualizimin e Rrjeteve Sociale
- Llogarit automatikisht metrika të rrjetit si Centrality, Density, Groupings
- I përdorur në hulumtime për komunikim, siguri, analiza sociale

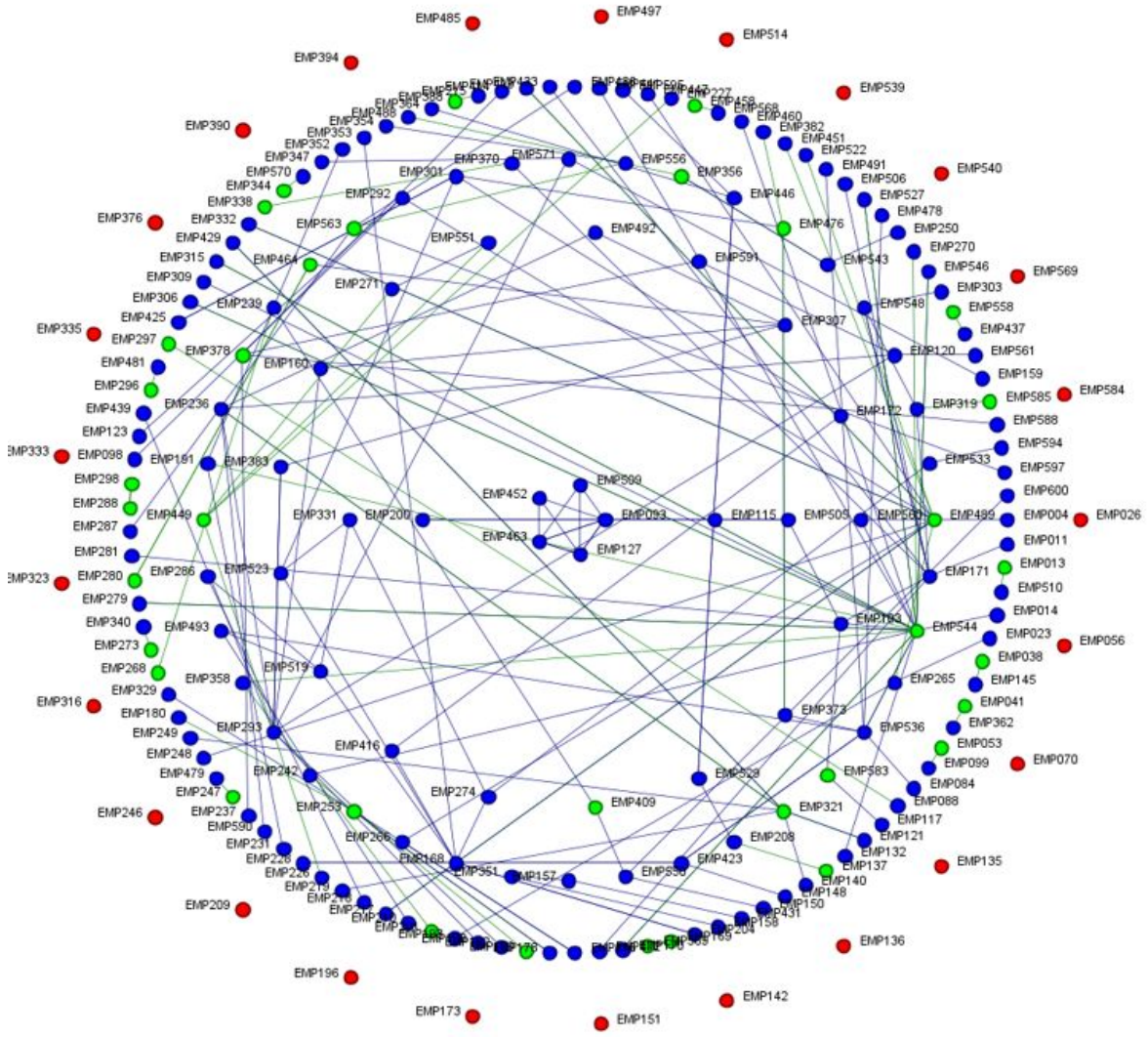






Analiza e Grafeve - Vizualizimi i Grafit

- Vizualizimi i rrjetit EmailNetwork
- Përdorimi i dataset-it Email Dataset Enron
- Dizajnimi përmes layout-it
- Përdorimi i atributit Node apperance



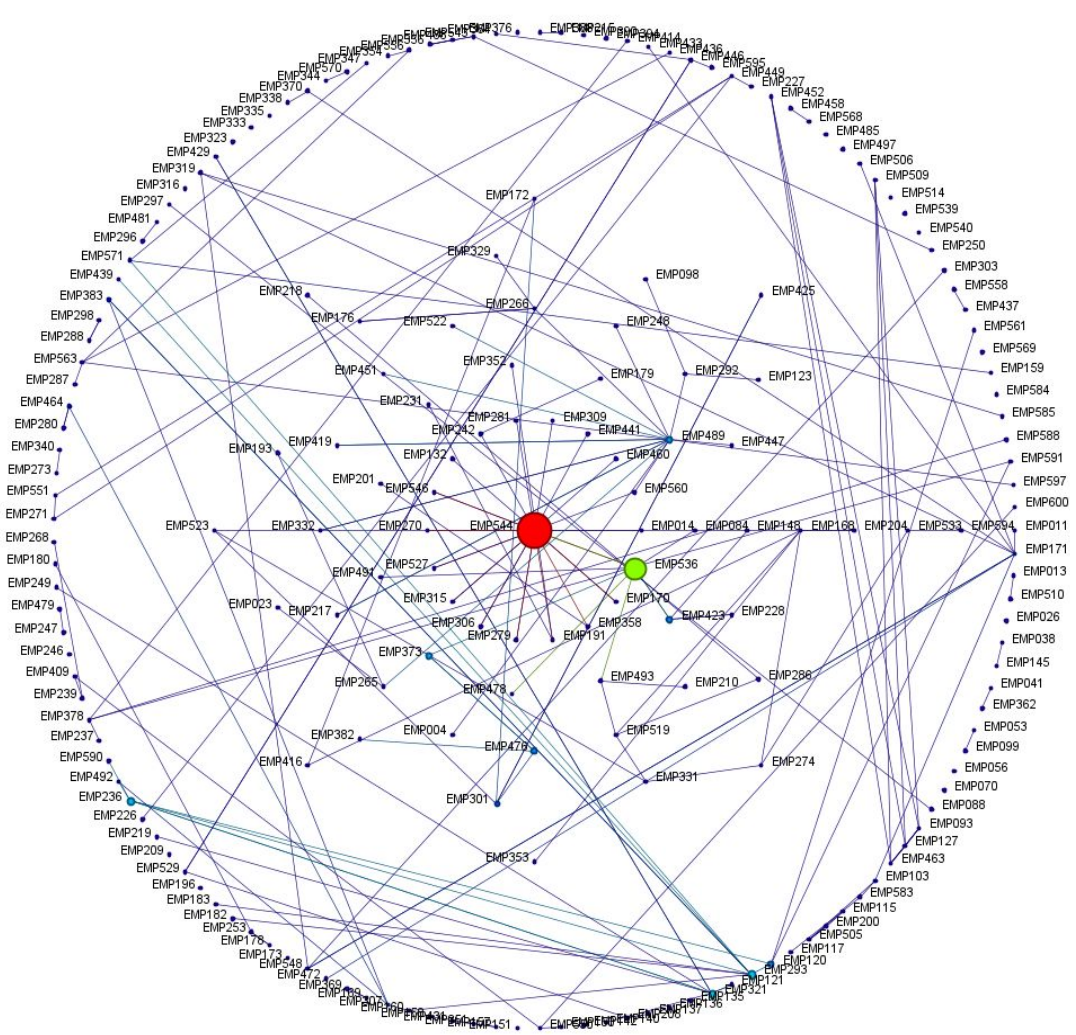
Analiza e Grafeve - Kalkulimi i metrikave

- Degree Centrality
- Betweenness
- Closeness

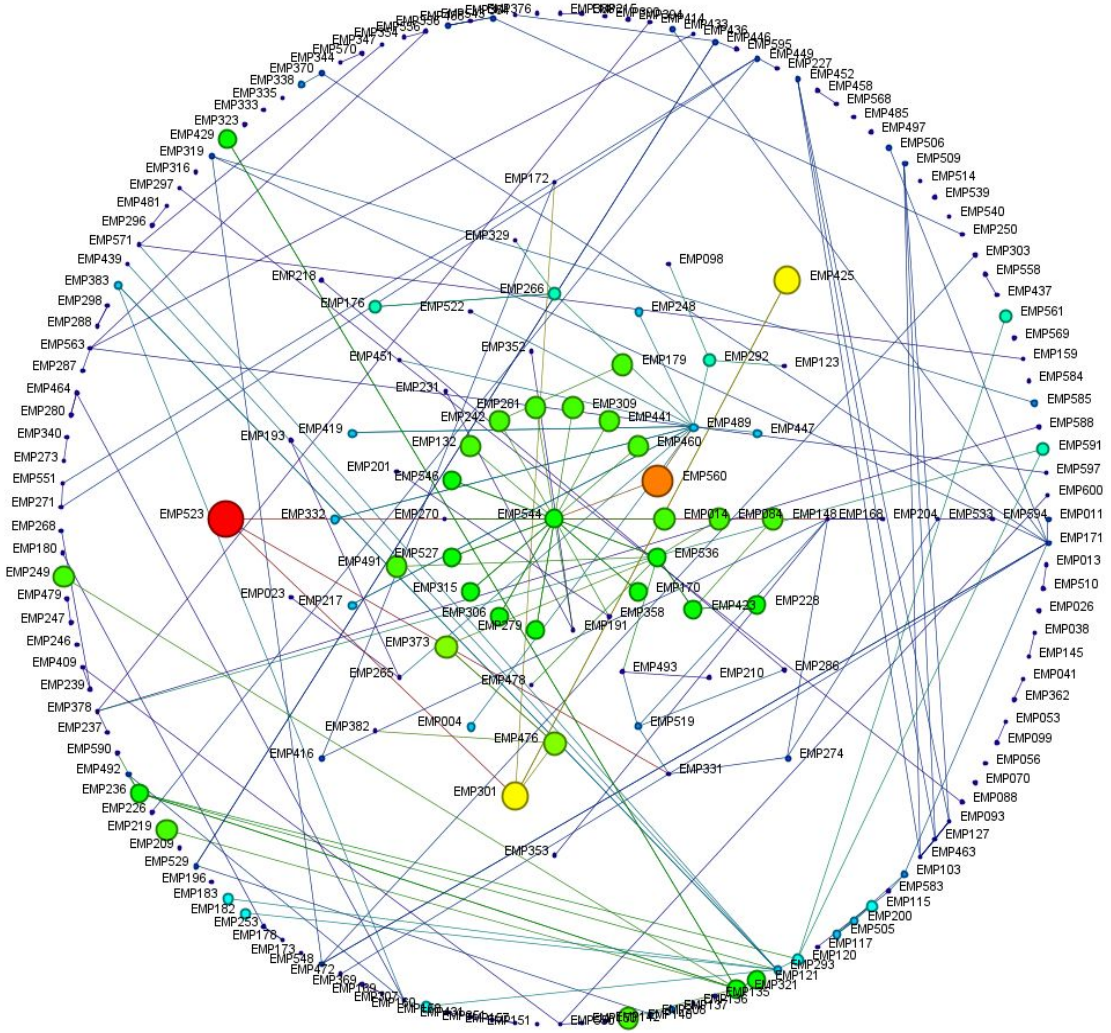
Node	Degree Centrality	Closeness Centrality	Betweenness Centrality
EMP544	0.017	0.000135	0.007
EMP536	0.006	0.000135	0.005
EMP171	0.005	0.000128	0.000732
EMP519	0.004	0.000129	0.0
EMP548	0.004	0.000128	0.000293
EMP489	0.003	0.00013	0.001
EMP168	0.003	0.000126	0.000122
EMP293	0.003	0.00013	0.002
EMP505	0.003	0.00013	0.000293
EMP200	0.003	0.00013	0.0
EMP321	0.002	0.000135	0.002
EMP423	0.002	0.000135	0.001
EMP084	0.002	0.000136	0.0
EMP014	0.002	0.000136	0.0
EMP236	0.002	0.000135	0.002
EMP476	0.000695	0.000137	0.001
EMP373	0.000695	0.000137	0.001
EMP120	0.000442	0.000131	0.001
EMP301	0.000442	0.000139	0.000951
EMP560	0.000379	0.000141	0.0
EMP132	0.00019	0.000136	0.0
EMP523	0.00019	0.000143	0.0
EMP425	0.000126	0.000139	0.0
EMP140	6.32e-05	0.000136	0.0

Analiza e Grafeve - Metrikat përmes vizualizimeve

- Betweenness



- Closeness



Analiza e Grafeve - Gephi

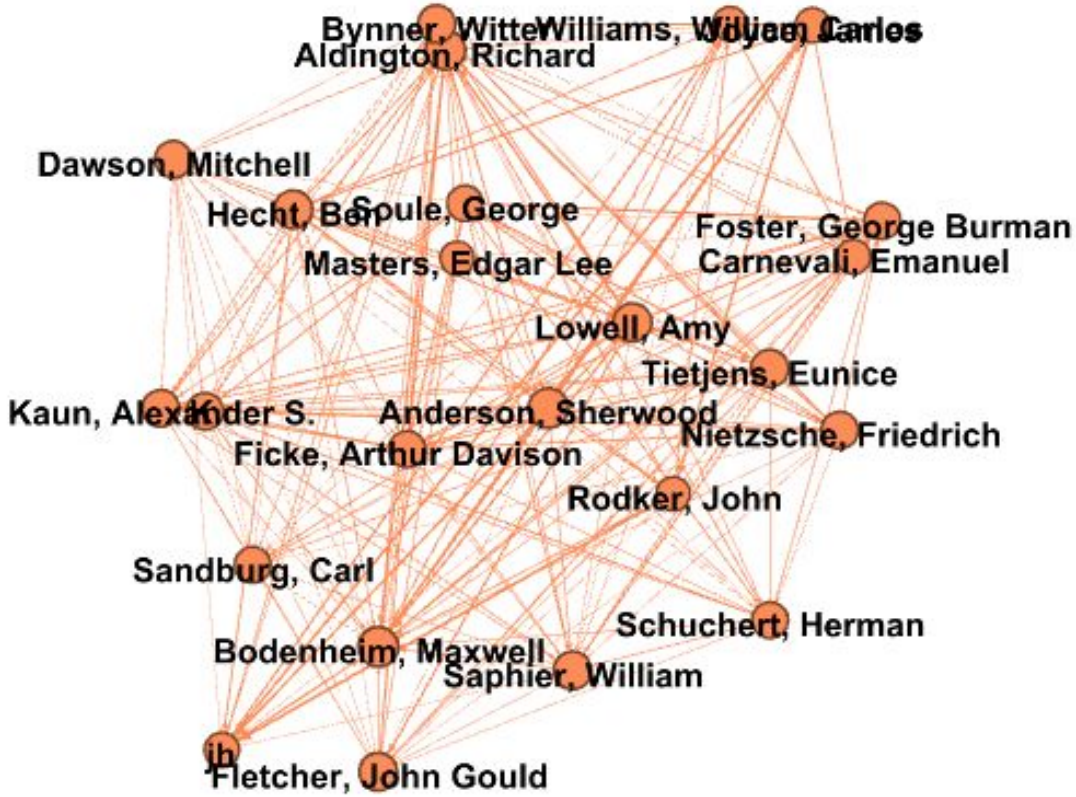
- Mjet për Analizën dhe Vizualizimin e Rrjeteve Sociale
- Real Time Preview
- Lehtë për përdorim vizual

Analiza e Grafeve - Gephi

Dataset: Poetry (1912-1922) and The Little Review (1914-1922)

nodes.csv & edges.csv

Nodes: 391
Edges: 2,426



Analiza e Grafeve - Gephi

- Degree Centrality
- Betweenness
- Closeness

Id	Degree	Closeness Centrality	Betweenness Centrality
Williams, William Carlos	55	1.0	0.090909
Tietjens, Eunice	106	0.866667	2.130592
Soule, George	55	0.73913	0.0
Schuchert, Herman	41	0.727273	0.071429
Saphier, William	50	0.833333	5.189683
Sandburg, Carl	47	0.692308	2.392496
Rodker, John	45	0.64	4.090909
Nietzsche, Friedrich	48	0.818182	0.404762
Masters, Edgar Lee	36	0.0	0.0
Lowell, Amy	101	0.875	0.404762
Kaun, Alexander S.	54	0.727273	3.033333
K.	48	0.75	1.1
Joyce, James	57	0.607143	1.416306
jh	78	0.777778	1.082973
Hecht, Ben	59	0.884615	0.0
Foster, George Burman	56	1.0	2.011905
Fletcher, John Gould	67	0.833333	9.847258
Ficke, Arthur Davison	61	0.777778	1.511905
Dawson, Mitchell	36	0.8125	3.385714
Carnevali, Emanuel	64	0.0	0.0
Bynner, Witter	50	0.833333	4.130592
Bodenheim, Maxwell	85	1.0	10.37583
Anderson, Sherwood	83	1.0	0.495671
Aldington, Richard	94	1.0	13.832973

Analiza e Grafeve - Gephi

- William Carlos Williams dhe Amy Lowell ka Degree më të lartë - shumë lidhje me autorë të tjerë, bashkëpunuese.
- Richard Aldington & Fletcher kanë Betweenness të lartë - janë "urë" mes grupeve.
- Bodenheim ka Closeness 1.0 - pozicion qendror, afër me të gjithë.
- Rrjeti tregon një komunitet shumë të lidhur të autorëve që kanë bashkëpunuar në revistat Poetry dhe The Little Review gjatë viteve 1912–1922.

Faleminderit për Vëmendjen!