



SISTEMI E ARCHITETTURE PER BIG DATA
PROGETTO #1

Processamento Distribuito di Big Data

Con Apache Hadoop e Apache Spark

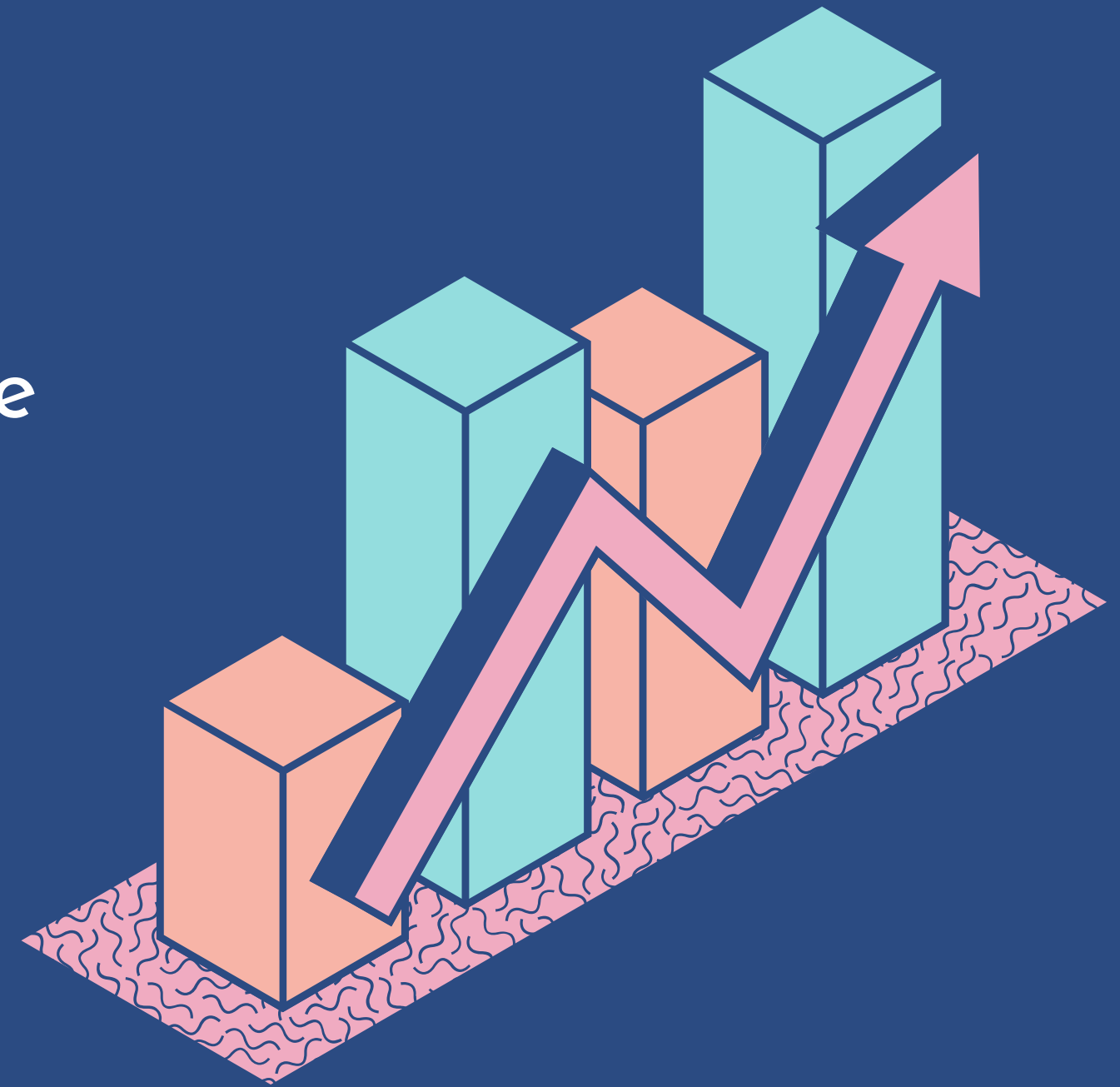
Lo scopo del progetto

Cosa

Fornire un'analisi quantitativa sull'andamento della campagna vaccinale italiana sul COVID19

Come

Tramite un'architettura di batch-processing sviluppata *ad-hoc* con Apache Hadoop e Apache Spark

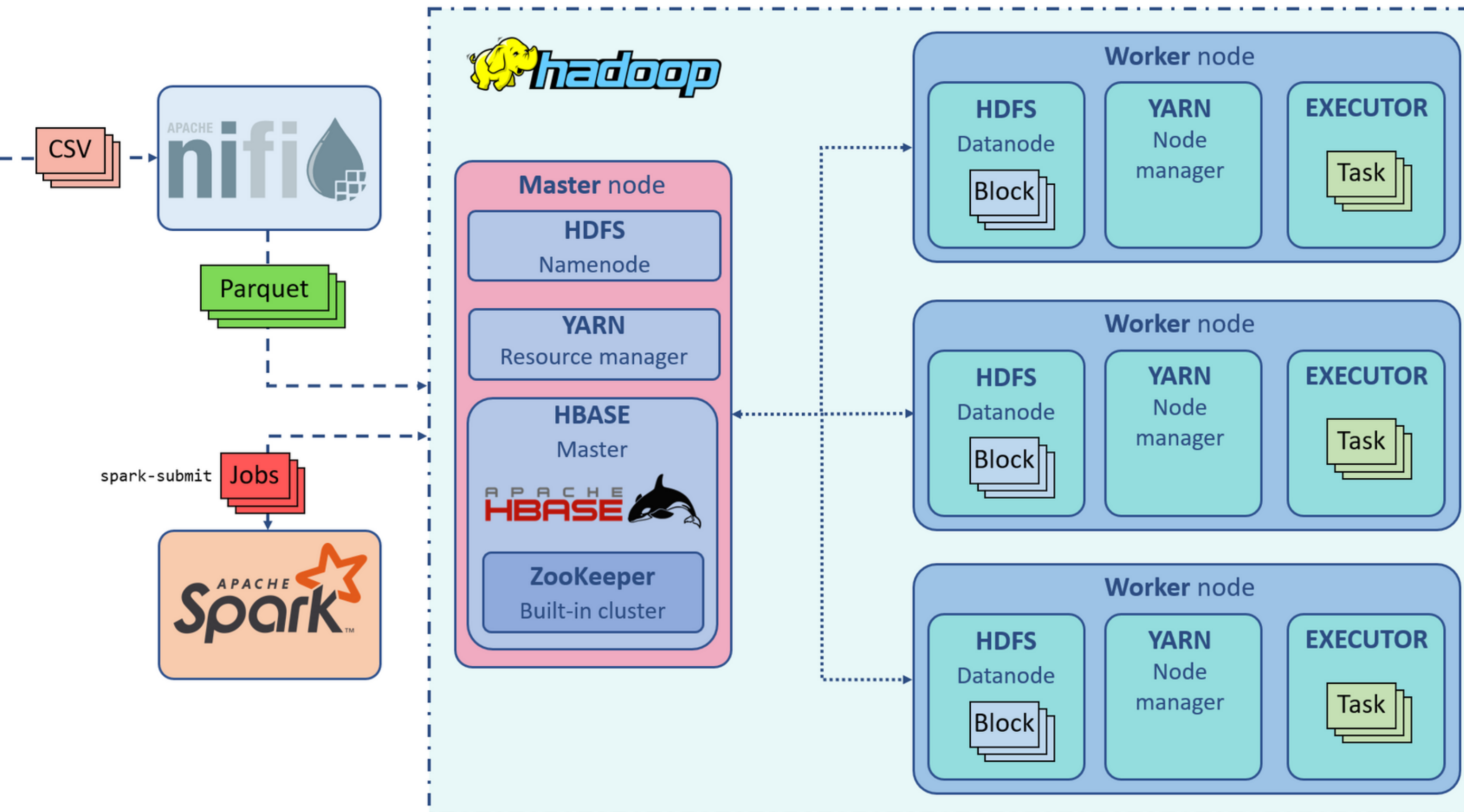


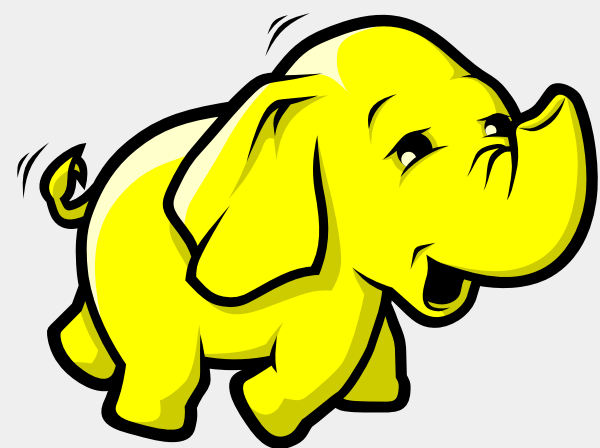
Architettura



As a whole

Schematizzazione
dell'architettura





Apache Hadoop

HDFS

- Storage Master/Worker
- Distribuito
- Dati replicati
- Throughput elevato
- Fault tolerance
- Utilizzato come:
 - Data source
 - Output destination

YARN

- Architettura Master/Worker
- Resource management
- Job scheduling
- Task eseguiti in parallelo
- Località dei dati (0 latenza)
- Utilizzato come:
 - Resource manager di Spark

HBASE

- Key-value & column-based
- Singolo Master/Region server
- Real-time random access
- Si appoggia su HDFS
- Utilizzato come:
 - Output destination



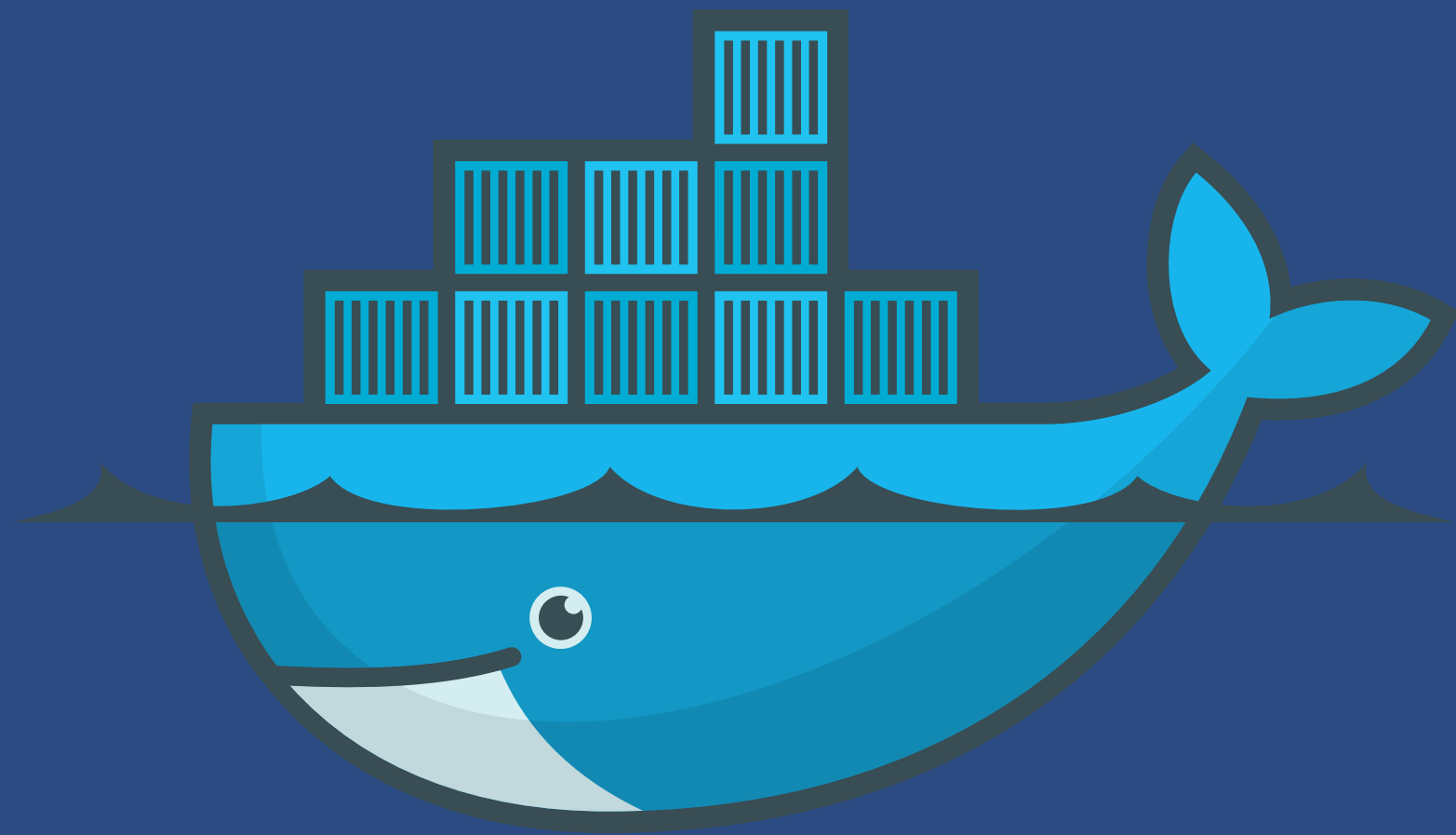
Punti chiave

- Processamento distribuito
- Map-reduce
- Performance superiori ad Apache MapReduce
- In-Memory caching
- SparkSQL e Spark MLlib

Uso

- Spark Core per Query 1, 2 e 3
- Spark SQL per Query 1 e 2
- Spark MLlib per Query 3

Deployment



DOCKER CONTAINER


- Simulazione di nodi della rete distinti
- Set-up e Clean-up automatico dei container
- Latenza pari a 0
- Risorse computazionali non ottimali
- Approfondimento nella configurazione

Nifi data ingestion



InvokeHTTP





InvokeHTTPSomministrazioneVaccin...

InvokeHTTP 1.13.2

org.apache.nifi - nifi-standard-nar


In	0 (0 bytes)	5 min
Read/Write	0 bytes / 0 bytes	5 min
Out	0 (0 bytes)	5 min
Tasks/Time	0 / 00:00:00.000	5 min

Permette di scaricare un file
attraverso una richiesta di GET
all'apposito URL

Property		Value
HTTP Method	?	GET
Remote URL	?	https://raw.githubusercontent.com/italia/covi...
SSL Context Service	?	No value set
Connection Timeout	?	5 secs
Read Timeout	?	15 secs
Idle Timeout	?	5 mins
Max Idle Connections	?	5
Include Date Header	?	True
Follow Redirects	?	True
Disable HTTP/2	?	False
Attributes to Send	?	No value set
Useragent	?	No value set

UpdateAttribute






UpdateAttributeSVSL

UpdateAttribute 1.13.2

org.apache.nifi - nifi-update-attribute-nar

In	0 (0 bytes)	5 min
Read/Write	0 bytes / 0 bytes	5 min
Out	0 (0 bytes)	5 min
Tasks/Time	0 / 00:00:00.000	5 min

Permette di modificare il nome del file attraverso l'aggiunta di una proprieta personalizzata (filename)

Property		Value	
Delete Attributes Expression	?	No value set	
Store State	?	Do not store state	
Stateful Variables Initial Value	?	No value set	
Cache Value Lookup Cache Size	?	100	
filename	?	SomministrazioneVacciniSummaryLatest.parq...	

ReplaceText


Permette di rimuovere delle specifiche colonne dal file attraverso l'utilizzo di apposite espressioni regolari (search value , replacement value)

 <div> <div>ReplaceText</div> <div>ReplaceText 1.13.2</div> <div>org.apache.nifi - nifi-standard-nar</div> </div>			
In	0 (0 bytes)		5 min
Read/Write	0 bytes / 0 bytes		5 min
Out	0 (0 bytes)		5 min
Tasks/Time	0 / 00:00:00.000		5 min

Property		Value
Search Value	?	^((?:.*,){3})(?:.*,){7})(.*)
Replacement Value	?	\$1
Character Set	?	UTF-8
Maximum Buffer Size	?	1 GB
Replacement Strategy	?	Regex Replace
Evaluation Mode	?	Line-by-Line
Line-by-Line Evaluation Mode	?	All

ConvertRecord





ConvertRecord

ConvertRecord 1.13.2

org.apache.nifi - nifi-standard-nar

In	0 (0 bytes)	5 min
Read/Write	0 bytes / 0 bytes	5 min
Out	0 (0 bytes)	5 min
Tasks/Time	0 / 00:00:00.000	5 min

Permette di trasformare il formato di un file : nel nostro caso la conversione avviene da csv a parquet

Ciò è possibile grazie all'uso di 2 Controller Service :

- CSVReader
- ParquetRecordSetWriter

Property		Value	
Record Reader	?	CSVReader	→
Record Writer	?	ParquetRecordSetWriter	→
Include Zero Record FlowFiles	?	true	

	CSVReader	CSVReader 1.13.2	org.apache.nifi - nifi-record-serializa...	 Enabled	NiFi Flow	 
	ParquetRecordSetWriter	ParquetRecordSetWriter 1.13.2	org.apache.nifi - nifi-parquet-nar	 Enabled	NiFi Flow	 

PutHDFS



	 PutHDFS PutHDFS 1.13.2 org.apache.nifi - nifi-hadoop-nar
In	0 (0 bytes) 5 min
Read/Write	0 bytes / 0 bytes 5 min
Out	0 (0 bytes) 5 min
Tasks/Time	0 / 00:00:00.000 5 min

Permette la scrittura di files su l' hdfs
specificato

Property		Value
Hadoop Configuration Resources	?	/opt/nifi/core-site.xml,/opt/nifi/hdfs-site.xml
Kerberos Credentials Service	?	No value set
Kerberos Principal	?	No value set
Kerberos Keytab	?	No value set
Kerberos Password	?	No value set
Kerberos Relogin Period	?	4 hours
Additional Classpath Resources	?	No value set
Directory	?	hdfs://master:54310/files
Conflict Resolution Strategy	?	ignore
Block Size	?	No value set
IO Buffer Size	?	No value set
Replication	?	No value set

Queries

PST = file Punti Somministrazione Tipologia

SVSL = file Somministrazione Vaccini Summary Latest

SVL = file Somministrazione Vaccini Latest

TP = file Totale Popolazione

Query 1

PST

SVSL

MapToPair

(per associare ad ogni centro il valore "1")



ReduceByKey

(per contare il totale di centri per regione)



Join



MapToPair

(per dividere il totale dei vaccini per i centri della regione e per i giorni del mese : funzione "computeDailyDoses" per i giorni del mese)

ReduceByKey

(per sommare vaccini somministrati in una stessa regione in uno stesso mese)



MapToPair

(per isolare come chiave la sola sigla della regione)



Query 1 risultato

mese	area	valore_medio
gennaio-2021	Calabria	12.75
gennaio-2021	Abruzzo	16.61
gennaio-2021	Lombardia	85.43
gennaio-2021	Toscana	18.51
gennaio-2021	Basilicata	17.27
gennaio-2021	Campania	104.25
gennaio-2021	Piemonte	26.35
gennaio-2021	Liguria	11.67
gennaio-2021	ProvinciaAutonomaBolzano	23.27
gennaio-2021	FriuliVeneziaGiulia	25.05
gennaio-2021	Sicilia	28.57
gennaio-2021	EmiliaRomagna	40.61
gennaio-2021	Marche	23.56
gennaio-2021	Lazio	33.29

Query 2 (parte 1)

SVL

ReduceByKey

(per sommare vaccini somministrati da case farmaceutiche diverse nello stesso giorno e nella stessa regione per la stessa fascia di età)



MapToPair

(per isolare come chiave la coppia {sigla della regione, fascia di età})



GroupByKey

(per ottenere tutti i giorni di somministrazione e le somministrazioni effettuate per ogni chiave)



FlatMapToPair

(per estrapolare le sottoliste di giorni di uno stesso mese per ogni chiave : funzione "daysGroupedByMonth")



Filter

(per rimuovere i mesi aventi meno di 2 giorni di campagna vaccinale)



MapToPair

(per isolare come chiave la coppia {giorno mese successivo, fascia di età} ed applicare la regressione lineare)



Query 2 (parte 2)

SVL

GroupByKey

(per raggruppare le regioni e relative somministrazioni in base alla chiave {giorno mese successivo,fascia di età})

SortByKey

(per ordinare in base al mese ed alle fascia di età)

MapToPair

(per ordinare le classifiche ed estrapolare da ognuna la top 5 : funzione "iterableToListTop5")

Query 2 risultato

anno-fascia	area_1	previsione_1	area_2	previsione_2	area_3	previsione_3	area_4	previsione_4	area_5	previsione_5
2021-03-01 12-19	Puglia	6	Lazio	6	Sicilia	5	Veneto	5	Toscana	5
2021-03-01 20-29	Toscana	444	Puglia	363	Veneto	334	Piemonte	283	Lombardia	271
2021-03-01 30-39	Toscana	911	Campania	564	Puglia	545	Piemonte	393	Lombardia	388
2021-03-01 40-49	Toscana	1424	Campania	1150	Puglia	919	Lazio	880	Sicilia	678
2021-03-01 50-59	Campania	1867	Lazio	1159	Toscana	1069	Puglia	975	Sicilia	622
2021-03-01 60-69	Campania	960	Puglia	441	Lazio	286	Toscana	236	Veneto	163
2021-03-01 70-79	Lombardia	133	Lazio	91	Sardegna	54	Toscana	54	Sicilia	49
2021-03-01 80-89	Lazio	3192	EmiliaRomagna	3011	Lombardia	2986	Campania	2616	Veneto	2422
2021-03-01 90+	EmiliaRomagna	1436	Lombardia	1076	Lazio	705	Toscana	601	Piemonte	592
2021-04-01 12-19	Lombardia	30	Sicilia	25	Campania	22	Veneto	21	Lazio	20
2021-04-01 20-29	Lombardia	1234	Veneto	519	Lazio	291	Campania	278	Piemonte	255
2021-04-01 30-39	Lombardia	1906	Veneto	551	Lazio	537	Piemonte	344	Campania	323
2021-04-01 40-49	Lombardia	2618	Veneto	781	Lazio	615	Sardegna	488	Calabria	407
2021-04-01 50-59	Lombardia	3059	Veneto	1154	Lazio	860	Liguria	640	Sardegna	636
2021-04-01 60-69	Lazio	1544	Lombardia	1471	Veneto	1135	Sicilia	968	Campania	931
2021-04-01 70-79	Veneto	4268	Lazio	3899	EmiliaRomagna	3238	Sicilia	2951	Campania	2894
2021-04-01 80-89	Lombardia	11240	Veneto	7134	Piemonte	5468	EmiliaRomagna	5173	Lazio	4958

Query 3 (parte 1)

SVSL

GroupByKey

(per raggruppare i giorni di vaccinazione in base alla regione)



MapToPair

(per applicare la regressione per stimare per ogni regione il numero di vaccini al 1 giugno)



Union

(per aggiungere le predizioni per il giorno del 1 giugno all' rdd di partenza)



ReduceByKey

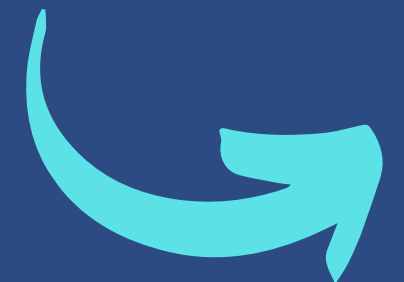
(per calcolare le vaccinazioni totali per regione)



Join



TP



Query 3 (parte 2)



MapToPair

(per calcolare la percentuale di popolazione vaccinata stimata
per ogni regione)



KMeans



BisectingKMeans

Query 3 risultato

KMeans

numero_cluster	WSSSE	[cluster]Regione
2	51.82	[0]ABR-[0]BAS-[0]CAM-[0]EMR-[0]FVG-[0]LAZ-[0]LIG-[0]LOM-[0]MAR-[0]MOL-[0]PAB-[0]PIE-[0]PUG-[0]TOS-[0]UMB-[0]VDA-[0]VEN-[1]CAL-[1]PAT-[1]SAR-[1]SIC
3	24.87	[0]ABR-[0]BAS-[0]CAM-[0]EMR-[0]FVG-[0]LAZ-[0]LOM-[0]MAR-[0]PAB-[0]PIE-[0]PUG-[0]TOS-[0]UMB-[0]VDA-[0]VEN-[1]LIG-[1]MOL-[2]CAL-[2]PAT-[2]SAR-[2]SIC
4	14.00	[0]ABR-[0]BAS-[0]CAM-[0]EMR-[0]FVG-[0]LAZ-[0]LOM-[0]MAR-[0]PAB-[0]PIE-[0]PUG-[0]UMB-[0]VDA-[1]PAT-[1]SIC-[2]LIG-[2]MOL-[3]CAL-[3]SAR-[3]TOS-[3]VEN
5	6.97	[0]ABR-[0]BAS-[0]EMR-[0]FVG-[0]LOM-[0]MAR-[0]PAB-[0]PUG-[0]UMB-[0]VDA-[1]CAL-[1]SAR-[2]PAT-[2]SIC-[3]LIG-[3]MOL-[4]CAM-[4]LAZ-[4]PIE-[4]TOS-[4]VEN

BisectingKMeans

numero_cluster	WSSSE	[cluster]Regione
2	58.04	[0]CAL-[0]CAM-[0]LAZ-[0]PAT-[0]PIE-[0]SAR-[0]SIC-[0]TOS-[0]VEN-[1]ABR-[1]BAS-[1]EMR-[1]FVG-[1]LIG-[1]LOM-[1]MAR-[1]MOL-[1]PAB-[1]PUG-[1]UMB-[1]VDA
3	41.32	[0]CAL-[0]CAM-[0]LAZ-[0]PAT-[0]PIE-[0]SAR-[0]SIC-[0]TOS-[0]VEN-[1]BAS-[1]EMR-[1]LOM-[1]MAR-[1]PAB-[1]PUG-[1]UMB-[1]VDA-[2]ABR-[2]FVG-[2]LIG-[2]MOL
4	16.03	[0]CAL-[0]PAT-[0]SAR-[0]SIC-[1]CAM-[1]LAZ-[1]PIE-[1]TOS-[1]VEN-[2]BAS-[2]EMR-[2]LOM-[2]MAR-[2]PAB-[2]PUG-[2]UMB-[2]VDA-[3]ABR-[3]FVG-[3]LIG-[3]MOL
5	14.84	[0]CAL-[0]PAT-[0]SAR-[0]SIC-[1]CAM-[1]LAZ-[1]PIE-[1]TOS-[1]VEN-[2]BAS-[2]EMR-[2]PAB-[2]VDA-[3]LOM-[3]MAR-[3]PUG-[3]UMB-[4]ABR-[4]FVG-[4]LIG-[4]MOL

Spark core VS Spark SQL



Confronto tempi Query 1



SparkSQL

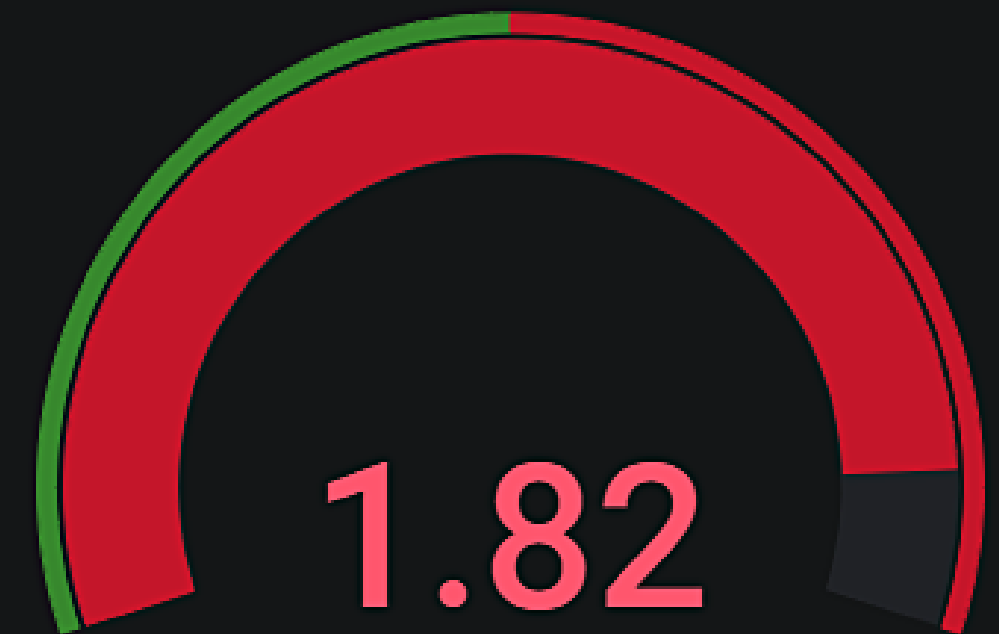


SparkCore

Confronto tempi Query 2



SparkSQL

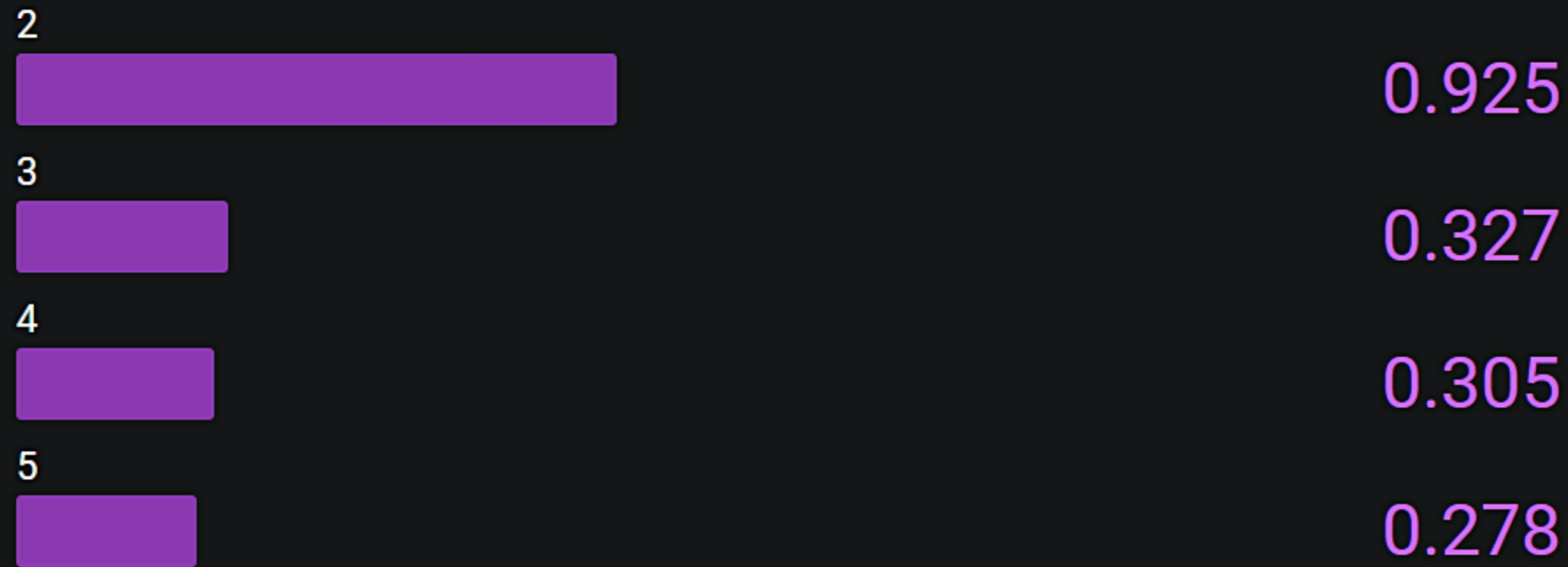


SparkCore

KMeans VS BisectingKMeans (Tempo)



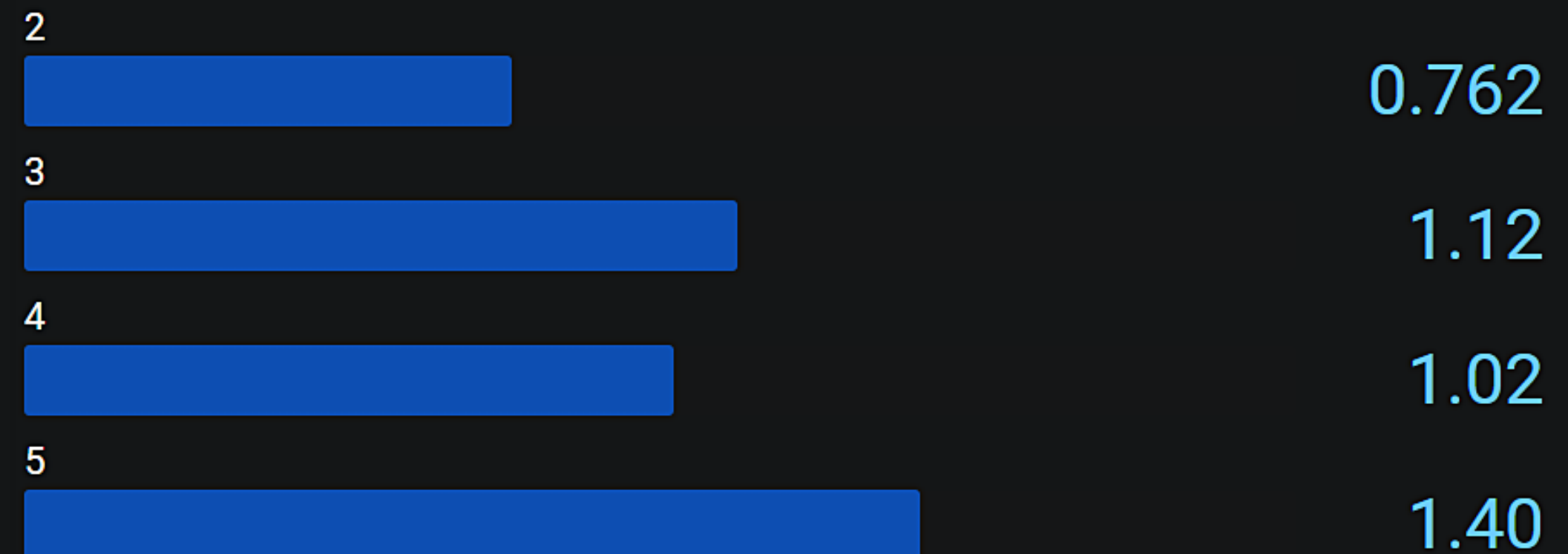
Tempi Kmeans al variare di k tra 2 e 5



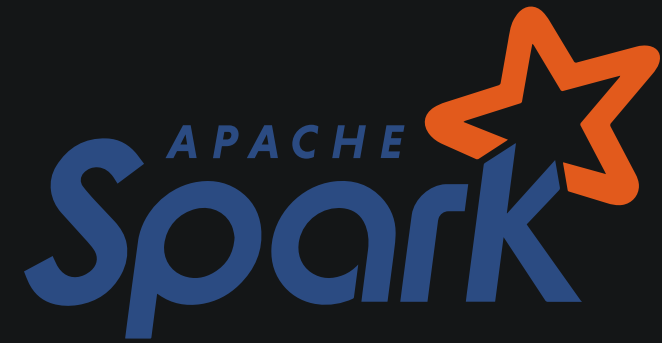
Tempo inversamente
proporzionale al numero di
clusters (KMeans)

Tempo direttamente
proporzionale al numero di
clusters (BisectingKMeans :
clustering gerarchico divisivo)

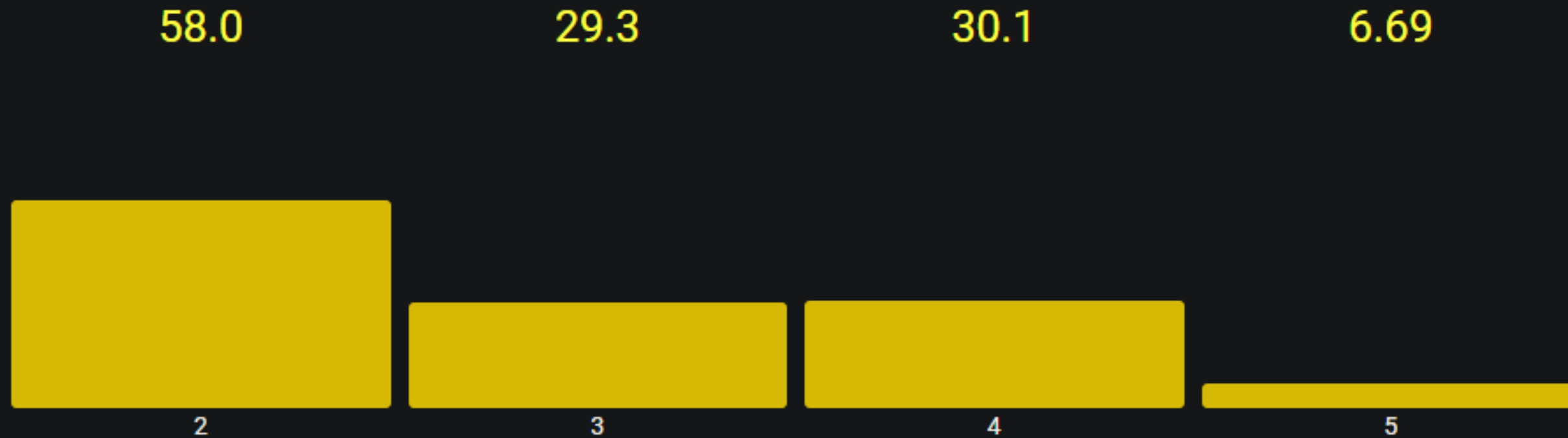
Tempi BisectingKmeans al variare di k tra 2 e 5



KMeans VS BisectingKMeans (WSSSE)

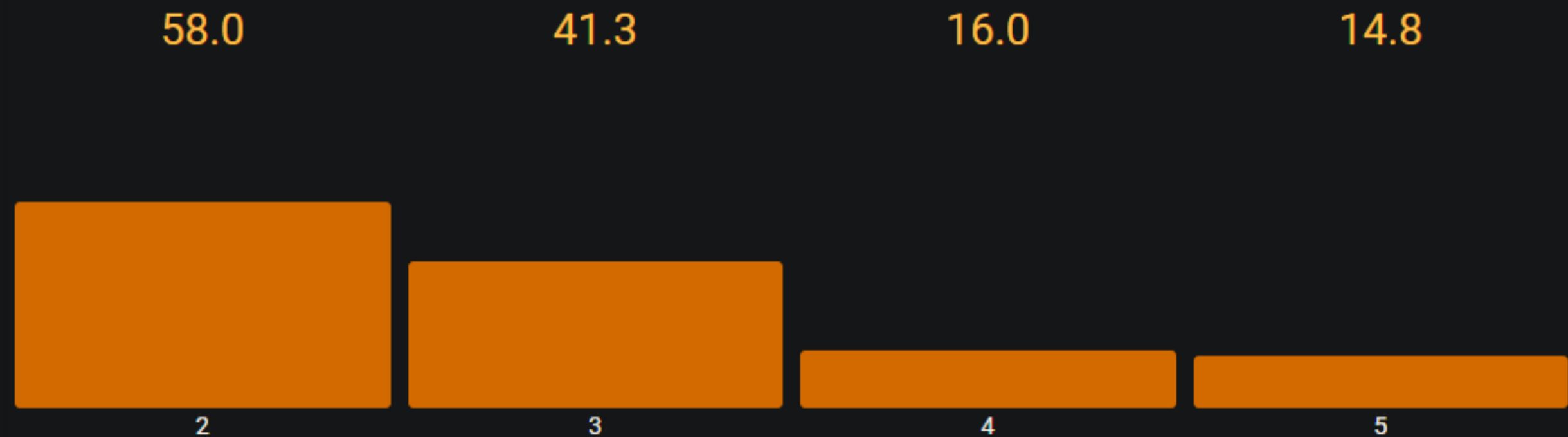


WSSSE Kmeans al variare di k tra 2 e 5



In entrambi i casi l'errore
decresce all'aumentare del
numero di clusters

WSSSE BisectingKmeans al variare di k tra 2 e 5



Il numero di clusters
probabilmente non è
ottimo poichè non si può
osservare l' "elbow" nel
grafico del WSSSE

Grafana dashboard



<https://snapshot.raintank.io/dashboard/snapshot/3EpFd4eW2fKRzxL7vKdqdM5b60Y0OEFi?orgId=2>

CREDITS

Andrea Paci

andrea.paci1998@gmail.com

Alessandro Amici

a.amici@outlook.it

Repository GitHub

<https://github.com/andreapaci/SABD>

Snapshot Grafana

<https://snapshot.raintank.io/dashboard/snapshot/3EpFd4eW2fKRzxL7vKdqdM5b60Y0OEFi>

