

Realizzazione e gestione di una nuova infrastruttura informatica al servizio della Pubblica Amministrazione denominata Polo Strategico Nazionale ("PSN"), di cui al comma 1 dell'articolo 33-septies del d.l. n. 179 del 2012

> CUP: J51B21005710007 CIG: 9066973ECE

# Manuale Utente PaaS Big Data

Data: 22/09/2023

PSN\_Manuale Utente PaaS Big Data

Ed. 1 - ver. 1.1



# QUESTA PAGINA È LASCIATA INTENZIONALMENTE BIANCA

# STATO DEL DOCUMENTO

TITOLO DEL DOCUMENTO			
Manuale Utente Secure Public Cloud su Cloud Provider Google			
EDIZ.	REV.	DATA	AGGIORNAMENTO
1	1.0	23/06/2023	Prima versione
1	1.1	22/09/2023	Aggiunto il Capitolo 6 "Data Governance"

NUMERO TOTALE PAGINE: 44	4

AUTORE:	
Team di lavoro PSN	Unità operativa Technology & Information

REVISIONE:	
Referente del Servizio	Paolo Trevisan

APPROVAZIONE:	
Direttore del Servizio	Antonio Garelli



## LISTA DI DISTRIBUZIONE

INTERNA A:

- Funzione Solution Development
- Funzione Technology Hub
- Funzione Sicurezza
- Referente Servizio
- Direttore Servizio

ESTERNA A:

• Direttore dell'Esecuzione Contrattuale (DEC) PSN

ing. Fabrizio Marchese



# INDICE

1	De	efinizioni e Acronimi	8
	1.1	DEFINIZIONI	8
	1.2		8
2	Ex	ecutive Summary	. 11
	2.1	SCOPO DEL DOCUMENTO	11
3	Dc	ata Lake	. 12
	3.1	GESTIONE BUCKET	12
	3.1.1	QUOTE	13
	3.1.2	VERSIONING	14
	3.1.3	LOCKING	14
	3.1.4	RETENTION	14
	3.1.5	Policy personalizzate	15
	3.2	GESTIONE FOLDERS E FILES	17
	3.3	Monitoring	19
4	Pro	ocessing	. 22
	4.1	Notebook	22
	4.1.1	Python	23
	4.1.1.1	Pyspark(Connect)	23
	4.1.1.2	Pyspark(Session)	24
	4.1.2	Scala	24
	4.1.3	SQL	25
	4.2	Scheduling	25
	4.3	Monitoring	26
5	Ev	ent Message	. 28
	5.1	Kafka UI	28



6

5.2	Gestione Topic	29
5.2.1	Creazione Topic	29
5.2.2	Modificare numero Partizione Topic	
5.2.3	Numero replica per Topic	31
5.2.4	RETENTION POLICY PER TOPIC	32
5.2.5	GENERAZIONE MESSAGGIO SU TOPIC	34
5.3	VISUALIZZAZIONE DATI	35
5.4	VISUALIZZAZIONE CONSOUMER GROUP	
5.5	UTILIZZO CLIENT PYTHON	
5.6	UTILIZZO CLIENT SPARK	37
D	ata Governance	
6.1	Data Lineage	
6.2	Data Lineage – Implementazione spark	41
6.3	Data Exploration	42



# LISTA DELLE TABELLE

Tabella 1: Glossario Definizioni	8
Tabella 2: Glossario Acronimi	10

# Definizioni e Acronimi

# *1.1* Definizioni

Definizione	Descrizione
PSN	È la nuova società che è stata costituita nell'ambito del progetto del Cloud Nazionale
ТВС	Il tema è stato discusso ma è in attesa di conferma dalle parti coinvolte
TBD	Il tema non è ancora stato discusso

Tabella 1: Glossario Definizioni

## 1.2 Acronimi

Acronimo	Descrizione
AD	Active Directory
APT	Advanced Persistent Threat
API	Application Program Interface
AV	AntiVirus
BaaS	Backup as a Service
CaaS	Container as a Service
CLI	Command Line Interface
CSP	Cloud Service Provider
DBE	DataBase Encryption
DDC	Data Discovery and Classification
DDoS	Distributed DoS
DE	Data Encryption
DLP	Data Loss Prevention
DM	Data Masking
DMZ	DeMilitarized Zone
DNS	Domain Name System
DoS	Denial of Service
DWDM	Dense Wavelength Division Multiplexing
EDE	Endpoint Disk Encryption
EDR	Endpoint Detection and Response
FIM	File Integrity Monitoring
FW	FireWall
Gbps	Gigabits per second
GUI	Graphical User Interface
HA	High Availability
HSM	Hardware Security Module



Acronimo	Descrizione
HTTP	HyperText Transfer Protocol
HTTPS	HTTP Secure
laaS	Infrastructure as a Service
IAG	Identity and Access Governance
I&AM	vedi IAM
IAM	Identity and Access Management
IDS	Intrusion Detection System
IP	Internet Protocol
IPS	Intrusion Prevention System
iSCSI	Internet SCSI
ISO	International Organization for Standardization
KMS	Key Management System
L2	Layer 2 (della pila ISO/OSI)
L3	Layer 3 (della pila ISO/OSI)
L4	Layer 4 (della pila ISO/OSI)
LAG	Link Aggregation Group
LAN	Local Area Network
LM	Log Management
LOM	Lights Out Management
MAC	Media Access Control
MC-LAG	Multi Chassis LAG
MDM	Mobile Device Management
MFA	Multi Factor Authentication
MPLS	MultiProtocol Label Switching
NAC	Network Access Control
NGFW	Next Generation FW
NL-SAS	Near Line SAS
NPB	Network Packet Broker
NTP	Network Time Protocol
OOB	Out of band
OSI	Open Systems Interconnection
PaaS	Platform as a Service
PA	Pubblica Amministrazione
PAM	Privileged Access Management
PdL	Postazione di Lavoro
PSN	Polo Strategico Nazionale
rpm	Rotation per minute
SaaS	Software as a Service
SAN	Storage Area Network
SAS	Serial Attached SCSI
SCSI	Small Computer System Interface
SEG	Security Email Gateway
SFP	Small Form-factor Pluggable
SFP+	Enhanced SFP
SIEM	Security Information and Event Management
SNMP	Simple Network Management Protocol
SOAR	Security Orchestration, Automation and Response



Acronimo	Descrizione
SOC	Security Operation Center
SQL	Structured Query Language
SR	Short Reach
SWG	Secure Web Gateway
ТВ	TeraByte
TBC	To Be Confirmed
TBD	To Be Defined
TI	Threat Intelligence and Infosharing
ToR	Top of Rack
VBR	Veeam Backup & Replication
VDOM	Virtual DOMain (Contesto Virtuale)
VLAN	Virtual LAN
VM	Vulnerability Management
VPN	Virtual Private Network
WAF	Web Application Firewall
WAN	Wide Area Network
XSS	Cross-Site Scripting

Tabella 2: Glossario Acronimi



## 2 Executive Summary

#### 2.1 Scopo del documento

Il documento ha lo scopo di fornire una guida all'utente finale delle funzionalità rilasciate nel servizio PaaS Big Data realizzato attraverso le soluzioni "Data Lake", "Processing", "Event Message" e "Data Governance".



## 3 Data Lake

Il componente Data Lake è basato sulla tecnologia MinIO. Lo scopo di tale componente è quello di implementare uno storage distribuito, supportato in lettura e/o scrittura da vari linguaggi di programmazione.

Per l'utilizzo all'interno della soluzione, è necessario effettuare il login sul Frontend del servizio "PaaS Big Data".



## **3.1** Gestione Bucket

Un bucket è un contenitore di dati del Data Lake al quale vengono applicate policy omogenee (es: accesso, retention, replicazione, crittografia, etc.). All'interno del bucket è possibile creare oggetti ai quali poter accedere con una semantica di tipo filesystem (struttura folders ad albero e files).



د ش Console		
	Buckets	
User	Search Buckets Q G Usage W Ubjects 7.2 <sub>MB</sub> 3	III 6 C Croate Bucket +
Administrator m Identity • Q Monitoring • A Notifications	data         Access: R/W           Created: 2023-06-08112:346:11Z         Access: R/W           Image: Created: 2023-06-08112:346:11Z         Boltects           63.0.ndt         1	Manage ⊕ Browce →
Tiers     Site Replication     Settings     Subscription	metastore         Access: RW           Created: 2023-06-0510941:21Z         Access: RW           Image: 0-Usage: 0-Directs         34,1xm           34,1xm         4	Manage ⊕ Browce →
<ul> <li>Ukonse</li> <li>Support</li> <li>SprOut</li> </ul>	pa-1-text-analytics-hf-keybert-multi           Created: 2023-06-2611055P312         Access: RWV           Otage         @ Objects           925.6 MB         14	Manage ⊕ Browce →

#### **3.1.1** Quote

Il Data Lake supporta la gestione di quote (occupazione massima volume di dati consentito) per bucket. Per modificare tale valore, è necessario seguire i seguenti step:

- 1. Aprire la pagina Data Lake  $\rightarrow$  Console
- 2. Cliccare sul bottone "Buckets"
- 3. Selezionare il bucket "testplan"
- 4. Cliccare su icona "edit Quota" e abilitare l'opzione settando una quota max di 10Mi

testp Access	l <b>an</b> : Custom		Delete Bucket	T Refr
Summary		Summary		
🔯 Enable Buck	et Quota		×	Reported
Enabled Quota*	10		OFF ON	<b>Usage:</b> 26.8 KiB
		Canc	cel Save	
ACC633		Versioning		1000
Anonymous		Current Status: Unversioned (Default)		

5. Cliccare su "Save" per rendere effettivo il settaggio



#### 3.1.2 Versioning

Il Data Lake supporta la gestione delle versioni degli oggetti. Per attivare tale funzionalità bisogna eseguire i seguenti step:

- 1. Aprire la pagina Data Lake  $\rightarrow$  Console
- 2. Cliccare sul bottone "Buckets"
- 3. Cliccare sul bottone "Create Bucket"
- 4. Creare un bucket "testret", selezionare "Versioning" e, opzionalmente, "Retention" in modalità "Compliance" per "30 days"
- 5. Cliccare "Create Bucket"
- 6. Caricare un file nel bucket appena creato

#### 3.1.3 Locking

Il Data Lake supporta l'object Locking degli oggetti, configurabile come segue:

- 1. Aprire la pagina Data Lake  $\rightarrow$  Console
- 2. Cliccare sul bottone "Buckets"
- 3. Cliccare sul bottone "Create Bucket"
- 4. Creare un bucket "testret", selezionare "Object Locking"
- 5. Cliccare "Create Bucket"
- 6. Caricare un file nel bucket appena creato

#### 3.1.4 Retention

Il Data Lake supporta la gestione del ciclo di vita degli oggetti, compreso il periodo di retention degli stessi, configurabile attraverso i seguenti step:

- 1. Aprire la pagina Data Lake  $\rightarrow$  Console
- 2. Cliccare sul bottone "Buckets"
- 3. Cliccare sul bottone "Create Bucket"
- 4. Creare un bucket "testret", selezionare "Object Locking" e "Retention" in modalità "Compliance" per "30 days"



Bucket Name*	testret	
View Bucket Naming Rule	25	
Versioning OFF ON		
Object Locking		
Quota		OFF O
Retention		
Mode	Compliance	Governance
Validity*	30	day

#### 3.1.5 Policy personalizzate

Il Data Lake supporta la definizione di ACL con granularità a livello di bucket o di singolo oggetto. Il test seguente mostra la possibilità di definire tali ACL utilizzando la UI. E' possibile definire ACL avanzate utilizzando descrittori JSON nel formato compatibile AWS (rif: <u>https://docs.aws.amazon.com/IAM/latest/UserGuide/reference\_policies.html</u>).

Un esempio di inserimento di una policy custom :

- 1. Aprire la pagina Data Lake  $\rightarrow$  Console
- 2. Cliccare sul bottone "Buckets"
- 3. Selezionare il bucket "testplan"
- 4. Cliccare su "Access Policy" e selezionare la voce "Custom"



ess Policy		2
Custom		•
2-10-17",		
	Canc	el Set
	Custom	Custom

5. Inserire la policy:

```
{
    "Version": "2012-10-17",
    "Statement": [
        {
             "Effect": "Deny",
"Principal": {
                  "AWS": [
                      "*"
                  ]
              },
              "Action": [
                 "s3:GetObject"
              ],
              "Resource": [
                  "arn:aws:s3:::testplan/testfolder/LICENSE"
             ]
        },
{
             "Effect": "Allow",
"Principal": {
"AWS": [
"*"
                  ]
              },
              "Action": [
                 "s3:GetObject"
              ],
              "Resource": [
                 "arn:aws:s3:::testplan/*"
             ]
        }
    ]
}
```



## 3.2 Gestione folders e files

All'interno dei bucket del nostro Data Lake sono memorizzati oggetti identificati all'interno del sistema da una semantica di tipo object storage (key-value). Il sistema consente tuttavia di poter manipolare logicamente questi oggetti utilizzando una semantica di tipo filesystem, ovvero operare su folders (con una struttura ad albero) e files.

Per navigare il filesystem è necessario:

• Aprire la pagina Data Lake  $\rightarrow$  Console

Nella finestra a destra, sarà possibile osservare tutti i bucket creati ed eventualmente, crearne di nuovi.

Per gestire il bucket che si vuole esplorare è necessario:

• Cliccare sul bottone "Object Browser" e selezionare il bucket designato

User		Filter Buckets			0
	Object Browser	1			۲
188	Access Keys	Name	Objects	Size	Access
	Description	🖶 jupyterlab	5 2	37.8 KiB	R/W
	Documentation	metastore	17	51.8 KiB	R/W
Admi	nistrator	spark-logs	121	9.9 MiB	R/W
E	Buckets	Testplan	0	0.0 B	R/W
0	Policies				

All'interno del bucket è possibile esplorare le folder esistenti o crearne di nuove come di seguito:

- 1. Cliccare sul bottone "Create new path"
- 2. Inserire il valore "testfolder" nella casella di testo
- 3. Cliccare sul bottone "Create"

Browser	tes Crea	stplan ated on: Mon. Jun 19 2023 09:16:27	Rewind 10	Refresh 🖒	Upload 🗘
entatio	:// Choose or c Current Path: /testplan New Folder Path*	reate a new path		×	e new path ://
S S		2	Clear	Create	1



Esattamente come per le folder, è possibile gestire i file in modalità Filesystem, una volta selezionata la folder, infatti, è possibile scaricare i file desiderati e/o caricarne di nuovi tramite il drag and drop o cliccando sul bottone "Upload".

otart typing to		2.55
🖌 🚟 🗸 📄 Test	Cerca	
Oggi		
		Create new 1 th 1/6
	LICENSE	
	Documento - 23 KB	
	Creato Giovedì 9 giugno 2022, 22:37	
	Annulla	

User	Object Browser	Filter Buckets				٩
(23)	Access Keys	Name	Objects	Size	Access	
		🗑 jupyterlab	5	37.8 KiB	R/W	
	Documentation	metastore	17	51.8 KiB	R/W	
Admi	inistrator	spark-logs	121	9.9 MiB	R/W	
E	Buckets	🗑 testplan	1	22.4 KiB	R/W	
â	Policies					



	TESTPIAN Created on: Mon, Jun 19 2023 09:16:27 (GMT+2) Access: PRIVATE 22.4 KiB - 1 Object	Rewind 9	Refresh 🖒 Upload 📫
<	testplan		Create new path ://
	▲ Name	Last Modified	Size
	testfolder		-
	Testfolder testplan Created on: Mon, Jun 19 2023 09:16:27 (GMT+2) Access: PRIVATE 22.4 KiB - 1 Object	Rewind 🔊	Refresh C Upload 1
<	testfolder testplan Created on: Mon, Jun 19 2023 09:16:27 (GMT+2) Access: PRIVATE 22.4 KiB - 1 Object testplan / testfolder	Rewind 9	Refresh C Upload 1
	testfolder testplan Created on: Mon, Jun 19 2023 09:16:27 (GMT+2) Access: PRIVATE 22.4 KiB - 1 Object testplan / testfolder Anme	Rewind 🔊	Refresh C Upload 1 Create new path :// Size

### 3.3 Monitoring

Per ogni Bucket infine, è possibile monitorare e modificare i vari aspetti (policy, percentuali utilizzo della memoria, fattore di replica ecc) tramite la pagina dedicata, raggiungibile attraverso i seguenti step:

- 1. Aprire la pagina Data Lake  $\rightarrow$  Console
- 2. Cliccare sul bottone "Buckets"





3. Cliccare, relativamente al Bucket desiderato, il tasto "Manage"

testpl	lan			Manage 😚	Browse →
Created:	2023-06-19T08:10:	10Z Access: R/W			Browse v
	<b>0</b>	() Objects		Manage Bucket	
	O Usage				
	<b>172.2</b> мів	4	•		



Access: Public		Delete I	Bucket 🗍 Refresh 🖒
Summary	Summary		
Events	Access Policy: 🖉	Encryption: <i>Disabled</i>	<b>Reported Usage:</b> 172.2 MiB
Replication	Replication: Disabled	Object Locking: <ul> <li>Disabled</li> </ul>	
Lifecycle	Tags:	Quota: 🖉	
Access Audit	+ Add tag Versioning		
Access Rules	Current Status: Unversioned (Default)		



## 4 Processing

Il modulo di processing è basato sull' utilizzo della libreria "Apache Spark", la quale permette di effettuare elaborazioni parallele e distribuite su più nodi.

PSN	×	O 237.18 pm 27 glugno 2023			e rossi leonardo.dev
"≓ Big Data େ Data Lake ⁰o Processing & Event Message		.vyeds: 532 Drives: 1 = 1.3390/3117 Gib	Events Name: local Broker Count: 1	Status: online	
'≡ Artificial Intelligence					
53 Semantic Search					
		3 12 Total, 4 Used 12 Zotal, 4 Used 12 Zotal, 6 Used 30627094951-172.19.0.9.49677 Used: 1 0 GB 00627094953-172.19.0.15-33427 Used: 1 0 GB 00627094953-172.19.0.16-42841 00627094953-172.19.0.16-42841 Used: 2 5 GB Used: 2 Ose GB	Applications App: Thrift JDBC/ODBC Server User: root Memory Executor: 2 048G/B State: RUNNING App: Spark Connect server User: root Memory Executor: 2 048G/B State: RUNNING	Cores: 2 Memory Stave: 2.048G/B Cores: 2 Memory Stave: 2.048G/B	

### 4.1 Notebook

Il modulo di processing mette a disposizione la possibilità di inizializzare dei notebook, con supporto a diversi linguaggi quali Python e Scala cliccando, dalla console, "Big Data" - >"Processing" -> "Notebook".



6	السابعة	🗆 Joh Manifaring 🎧	Liberatoria - El Schedur Pre	
Ce	ovac ay civy of	C soo manoning ( ) i		
0	File Edit View Run	Kernel Tabs Settings Hel		
	+ k ±		In Untited Lipynb	
	Filter files by name		H + X - > = C + Code C	) (Connect)
0				무 표
=	Name	- Last Modifie	field	
*				

#### **4.1.1** Python

Per utilizzare un notebook con kernel Python, basta entrare nella pagina di processing e selezionare un nuovo notebook Python 3(ipykernel).

Filter files by na	me ्			
/		Notebool	k	
Name 🔶	Last Modified			
bench_con	18 ore fa			
bench_sess	10 giorni fa			Ρ
📕 demo1.ipynb	17 giorni fa	Python 3	PySpark 340	PySpark 340
📕 demo2.ipynb	17 giorni fa	(ipykernel)	(Connect)	(Session)
📕 HL7.ipynb	17 giorni fa			
LeggiCSVC	18 giorni fa	>_ Console		
📕 spark-test.i	17 giorni fa			
📕 test.ipynb	17 giorni fa	_		
Untitled.ipy	10 giorni fa		P	P
		Python 3 (ipykernel)	PySpark 3.4.0 (Connect)	PySpark 3.4.0 (Session)
		S Other		

#### *4.1.1.1 Pyspark(Connect)*

Per utilizzare un notebook con kernel PySpark (Connect), basta entrare nella pagina di processing e selezionare un nuovo notebook PySpark (Connect).





#### 4.1.1.2 Pyspark(Session)

Per utilizzare un notebook con kernel PySpark (Session), basta entrare nella pagina di processing e selezionare un nuovo notebook PySpark (Session).

Filter files by nam	ıe ୦	
-/		Notebook
Name 🔶	Last Modified	
bench_con	18 ore fa	
📕 bench_sess	10 giorni fa	
🗖 demo1.ipynb	17 giorni fa	Python 3 PySpark 340 PySpark 340
🗖 demo2.ipynb	17 giorni fa	(ipykernel) (Connect) (Session)
📕 HL7.ipynb	17 giorni fa	
LeggiCSVC	18 giorni fa	Console
🗖 spark-test.i	17 giorni fa	
🗖 test.ipynb	17 giorni fa	
📕 Untitled.ipy	10 giorni fa	
		Python 3 PySpark 3.4.0 PySpark 3.4.0 (ipykernel) (Connect) (Session)
		\$ Other

## **4.1.2** Scala

Per utilizzare un notebook con kernel Spark - Scala, basta entrare nella pagina di processing e selezionare un nuovo notebook Scala.





## **4.1.3** SQL

Per utilizzare un notebook con kernel Spark - SQL, basta entrare nella pagina di processing e selezionare un nuovo notebook Spark-SQL.

e Edit View Run Kernel Tabs S	Settings Help
+ 🗈 🛨 C	E test-py.ipynb × ■ test-sc.ipynb × ☑ Launcher × ■ root@dcc9f063
Filter files by name Q	test
Name 🔺 Last Modified	Notebook
• 📃 test-py.ipy 15 minuti fa	
• 🖪 test-sc.ipy 1 minuto fa	Python 3 PyS PySpark (here and spark - Scala
	(ipykernel) Session
	Spark - SQL

#### 4.2 Scheduling

Il componente processing consente la creazione e/o la schedulazione di job attraverso i seguenti step:



- 1. Aprire la pagina Processing  $\rightarrow$  Scheduler Page
- 2. Modificare i Job nella "Lista Applicazioni" o andare sul tab Impostazioni -> Jobs

. KLEONARDO	2:17:19 pm 26 glugno 2023				😝 rossi.Jeonardo.dev 😟 Italiano
🖗 SQL 🛱 Livy UI 🖵 Job Monitoring	() Notebook 📴 Scheduler Page				
		Lit	sta Applicazioni		
Αρρ					Azioni
SparkPl	Transform and Load main database	High Performance	Alle 10:00	<b>0</b> 0	+ /
SparkPi Develop	Transform and Load main database	Low Performance			+ / 0
SparkPi (Python)	Transform and Load main database	Low Performance	Alle 00:00, solo il lunedi, mercoledi, e venerdi	0 0	+ /

#### 4.3 Monitoring

Il componente processing consente la creazione e/o la schedulazione di job attraverso i seguenti step:

- 1. Aprire la pagina Processing  $\rightarrow$  Job Monitoring
- 2. Cliccare sull'applicazione desiderata
- 3. Cliccare su stdout/stderr sui worker in esecuzione

@ SQL 冊 Livy UI □ Job Monitoring 〈〉 Not	ebook	🗊 Scheduler Page											
Sport 34.0 Spark Master at spark:	//spar	k-master:7077											
Uilt: spark-()spark-master:2077 Allew Yorken: 3 Correa In use: 17 Total, 4 Used Annorey In use: 170 GB Total, 40 OB Used Annorey In use: 170 GB Total, 40 OB Used Annorey In User, 40 OB Total, 40 OB Used Annorey In User, 40 OB Total, 40 OB Total, 40 Orivers: 0 Durahlp, 0 Completed Status: K./NC - Workers (3)													
Worker Id				Addre		St	ate	Cores	Memory			Resources	0
worker-20230623082306-172 18.0.9-46507 // 12 18.0.9-407 // 12 18.0.000 // 12 18.0.000 // 12 18.0.0000 // 12 18.0.0000 // 12 18.0.0000 // 12 18.0.0000 // 12 18.0.0000 // 12 18.0.0000 // 12 18.0.0000 // 12 18.0.0000 // 12 18.0.0000 // 12 18.0.0000 // 12 18.0.0000 // 12 18.0.00000							4.0 GiB (2.0 GiB Used)						
worker-20230623083207-172.18.0.14-35605				172.18	172.18.0.14:35605		IVE	4 (2 Used) 4.0 GIB (4.0 GIB Used)					
worker-20230623083208-172.18.0.15-43151	Increa-20230623083208-772.18.0.15-43151 ALIVE 4 (1 Used) 4.0 GB (2.0 GB Used)												
- Running Applications (2)													
Application ID		Name		Cores	Memory per Executor		Resources Pe	r Executor	Submitted Time	Use	State	5	Duration
app-20230623083213-0001	(icill)	Thrift JDBC/ODBC Server		2	2.0 GiB				2023/06/23 08:32:13 root		RUNN	ING	75.8 h
app-20230623083212-0000	(kill)	Spark Connect server		2	2.0 GIB				2023/06/23 08:32:12 root		RUNN	ING	75.8 h
- Completed Applications (1)													
Application ID	Name		Cores	Memory per	Executor	Resources	Per Executor		ubmitted Time	User	State		Duration
app-20230626121740-0002	SampleS	park.Job\$	0	4.0 GIB					023/06/26 12:17:40	root	FINISHED		1s



ID: app-202306230 Name: Thrift JDBC/ User: root Cores: Unlimited (2 Executor Limit: Un Executor Memory Executor Resource Submit Date: 2023 State: RUNNING Application Detail	83213-0001 ODBC Server granted) Iimited (2 granted) - Default Resource Profile: 2.0 GiB s - Default Resource Profile: /06/23 08:32:13 UI						
ExecutorID	Worker	Cores	Memory	Resource Profile Id	Resources	State	Logs
1	worker-20230623083207-172.18.0.14-35605	1	2048	0		RUNNING	stdout stderr
0	worker-20230623083208-172.18.0.15-43151	1	2048	0		RUNNING	stdout stderr



## 5 Event Message

Il modulo di "Event Message" è basato sull' utilizzo del message broker "Kafka", il quale permette di scambiare flussi di dati tra i vari producer/consumer.

psn X	O 237:18 pm 27 giugno 2023		🕃 rossi leonardo dev
'≡ Big Data ^			
🔓 Data Lake	Objects: 532 Drives: 1	Name: local Status: online Broker Count: 1	
•o Processing	je: 1.555673117 Gib		
🕲 Event Message			
'E Artificial Intelligence $\checkmark$			
53 Semantic Search			
	: 3 12 Total, 4 Used e: 12.288 GiB Total, 8.192 Gib Used	App: Thrift JDBC/ODBC Server User: root Cores: 2 Memory Executor: 2.048GiB Memory Slave: 2.048GiB	
	230627094951-172.19.0.9-43677 Used: 1	State: RUNNING App: Spark Connect server	
	6 GIB Used: 2.048 GIB	User: root Cores: 2 Memory Executor: 2.048GiB Memory Stave: 2.048GiB State: RUNNING	
	Used: 1 6 GiB Used: 2.048 GiB		
	230627094953-172.19.0.16-42841 Used: 2		
	6 GIB Used: 4.096 GIB		

#### 5.1 Kafka Ul

La web Application Kafka UI, fornisce un'interfaccia grafica per la gestione dei Topic. Un Topic può essere pubblicato e sottoscritto da varie applicazioni che possono adoperare come "Producer" e/o "Consumer", scrivendo e consumando flussi dati sul Topic. Tramite la Kafka UI è possibile non solo vedere i flussi di dati legati ad ogni Topic, bensì è possibile amministrare gli stessi (es. creazione, cancellazione, replica ecc...).



■   ※ LEO	NARDO © 2:45:54 pm 27 giugno 2023						erossil.leonardo.dev	Italia
Console								
Ul for Apache Kafka								o~ ∩ #
Dashboard	ashboard							
Brokers Topics Consumers	Onine) Offine I clusters O clusters							
hor	Only offline clusters     Cluster name							
	clare	3.4-IV0				0 Bytes	0 Bytes	

#### *5.2* Gestione Topic

La gestione dei Topic risulta essere un componente chiave del modulo "Event Message", di seguito, viene descritto come espletare le principali operazioni su essi.

#### *5.2.1* Creazione Topic

Per la creazione di nuovi Topic è necessario:

- 1. Aprire la pagina Event Message  $\rightarrow$  Web Ul
- 2. Cliccare su Topics  $\rightarrow$  Add a Topic
- 3. Impostare "Topic name" test001
- 4. "Number of partitions" 1
- 5. "Replication Factor" 1
- 6. Cliccare su "Create topic"



	Topics / Create
local • ^	
Brokers	Topic Name *
Topics	test001
Consumers	Number of partitions * Cleanup policy
ACL	1 Delete ~
	Min In Sync Replicats Replication Factor
	Min In Sync Replicas
	Time to retain data (in ms)
	Time to retain data (in ms)
	12 hours 1 day 2 days 7 days 4 weeks
	Max size on disk in GB Maximum message size in bytes
	Not Set V Maximum message size
	Custom parameters
	+ Add Custom Parameter

## *5.2.2* Modificare numero Partizione Topic

Per modificare il numero di partizioni di un Topic è necessario:

- 1. Aprire la pagina Event Message  $\rightarrow$  Web Ul
- 2. Cliccare su Topics e selezionare la topic "topic001" dall'elenco
- 3. Cliccare sull'icona a forma di tre puntini in alto a dx e selezionare "Edit settings" dall'elenco a tendina



verview Mes	view Messages Consumers Settings Statistics							peration has consequences.
Partitions 2	Replication Factor URP • 1 0		In Sync Replicas • Type 2 of 2 External		Type External		Clear messages Clearing messages is only allowed for with DELETE policy Recreate Topic Remove Topic	
Clean Up Policy DELETE				Message Count 0				
Partition ID	Replicas	First Offset		Next Offset		Messag	ge Count	
)	1	0		0		0		* *
	1	0		0		0		:

4. Incrementare a 2 il valore di "Number of partitions"

Danger Zone		
Change these parameters only if y	ou are absolutely sure what you are doing.	1
Number of partitions *		
2		\$ Submit
Replication Factor *	•	
1		Submit

5. Cliccare su "Submit" per rendere effettivo il settaggio

#### *5.2.3* Numero replica per Topic

Per modificare il numero di replica per un Topic è necessario:

- 1. Aprire la pagina Event Message  $\rightarrow$  Web Ul
- 2. Cliccare su Topics e selezionare la topic "topic001" dall'elenco
- 3. Cliccare sull'icona a forma di tre puntini in alto a dx e selezionare "Edit settings" dall'elenco a tendina



/erview Mes	sages Consumers	Settings Statistics	5		Edit settin Pay attenti especially	ngs on! This operation has important consequences.
Partitions 2	Replication Factor     URP •     In Sync Replicas •     Type       1     0     2 of 2     External		Type Clear message External Clearing message with DELETE poli			
Clean Up Policy DELETE			Message C O	punt	Recreate Remove 1	Торіс Горіс
artition ID	Replicas	First Offset	Next Offset		Message Count	
)	1	0	0		0	:
	1	0	0		0	:

4. Incrementare il valore "Replication Factor"

Danger Zone	
Change these parameters only if you are absolutely sure what you are doing.	
Number of partitions *	Submit
Replication Factor *	Submit

5. Cliccare su "Submit" per rendere effettivo il settaggio

#### *5.2.4* Retention policy per Topic

Per modificare la politica di retention di un Topic è necessario:

- 1. Aprire la pagina Event Message  $\rightarrow$  Web Ul
- 2. Cliccare su Topics e selezionare la topic "topic001" dall'elenco
- 3. Cliccare sull'icona a forma di tre puntini in alto a dx e selezionare "Edit settings" dall'elenco a tendina



verview Mes	ssages Consumers	Settings Statistics	5		Edit settings Pay attention! especially imp	This operation has ortant consequences.
Partitions 2	Replication Factor 1	URP • O	In Sync Replicas • 2 of 2	Type External	Clear messa Clearing mess with DELETE p	iges ages is only allowed for topic policy
Clean Up Policy DELETE			Messag O	e Count	Recreate To Remove Top	pic vic
artition ID	Replicas	First Offset	Next Of	fset	Message Count	
1	1	0	0		0	* * *
	1	0	0		0	:

4. Portare a "86400000 ms" (1d) il valore di "Time to retain data (in ms)"

1					
ime to reta	in data (in ms) 1d				
86400000					
12 hours	1 day 2 days 7 days 4 weeks				
/lax size on	disk in GB Maximum message size in bytes				
Max size on Not Set	disk in GB Maximum message size in bytes          V       1048588				
Max size on Not Set Custom pa	disk in GB Maximum message size in bytes 1048588 arameters				
Max size on Not Set Custom pa + Add (	disk in GB Maximum message size in bytes 1048588 arameters Custom Parameter				



Dashboard	Topics / test001	
Dev • ^	Overview Messages Consumers Settings Statistic	cs
Topics	Key	Value
Consumers	compression.type	producer
ACL	leader.replication.throttled.replicas	
	min.insync.replicas	1
	message.downconversion.enable	true
	segment.jitter.ms	0
	cleanup.policy	delete
	flush.ms	9223372036854775807
	follower.replication.throttled.replicas	
	Segmentarytes	1075741024
	retention.ms	86400000
	fiusinmessages	0220072000004770807
	message.format.version	3.0-IV1
	max compaction lag ms	0222272026854775807

## *5.2.5* Generazione messaggio su Topic

Per generare un messaggio su un Topic è necessario:

- 1. Aprire la pagina Event Message  $\rightarrow$  Web Ul
- 2. Cliccare su Topics e selezionare la topic "topic001" dall'elenco
- 3. Cliccare sul bottone "Produce Message" in alto a dx

Topics / tes	st001					Produce Message	:
Overview	Messages Consumers	Settings Statistics					
Partitions 2	Replication Factor 2	URP • O	In Sync Replicas • 4 of 4	Type Extern	Segment Size 0 Bytes	Segment Count 4	
Clean Up Poli DELETE	icy		Message o 0	unt			
Partition ID	Replicas	First Offset	Next Offset		Message Count		
0	1, 3	0	0		0	•	

- 4. Inserire "001" come key e "test" come message
- 5. Cliccare su "Produce Message" per confermare l'invio



	Value Serde	
String	<ul> <li>✓ String</li> </ul>	`
Keep contents		
Key		
T 001		
Value		
i cost		

### *5.3* Visualizzazione Dati

Per visualizzare i dati tramite Kafka UI è necessario cliccare su *"Topics"* e aprire la tab *"Messages"* del Topic di cui si vuol ispezionare il flusso di dati.



Overview	Messages Consur	ners Settings Statistics					
Seek Type	Offset	Partitions	Key Serde	Value Serde	~	Clear all	
Submit Q Search	⊗ +	Add Filters					Oldest
				DONE	<b>③</b> 7 ms	↓ 9 Bytes	a 2 messag
Offset	Partition Timestan	np Key Preview		Value Prev	view		

#### *5.4* Visualizzazione Consoumer Group

Kafka UI del modulo Event Message consente di monitorare i consumer attualmente attivi cliccando su "Consumers".

Dashboard	Consumers					
ocal • ^	Q Search by (	Consumer Group ID	8			
Topics	Group ID	Num Of Members	Num Of Topics	Messages Behind	Coordinator	State
Consumers	cg001	1	1	0	1	STABLE

#### 5.5 Utilizzo client Python

Il modulo Event Message essendo basato su Apache Kafka consente l'utilizzo delle librerie disponibili per i linguaggi di programmazione più diffusi. I seguenti step permettono l'utilizzo tramite il client Python.

- 1. Aprire la pagina Processing  $\rightarrow$  Notebooks
- 2. Creare un nuovo Notebook con kernel Python
- 3. Inserire il seguente codice in una cella:



```
# message value and key are raw bytes -- decode if necessary!
# e.g., for unicode: `message.value.decode('utf-8')`
print ("%s:%d:%d: key=%s value=%s" % (message.topic,
message.partition,
message.offset,
message.key,
message.value))
```

- 4. Mandare in esecuzione
- 5. Pubblicare un messaggio sul topic tramite web UI (vedere test precedenti)

#### **5.6** Utilizzo client Spark

Il modulo Event Message essendo basato su Apache Kafka consente l'utilizzo delle librerie disponibili per i linguaggi di programmazione più diffusi ed i framework di stream processing più diffusi. I seguenti step permettono l'utilizzo del modulo Event Manager tramite Apache Spark per il processing dei messaggi.

- 1. Aprire la pagina Processing  $\rightarrow$  Notebooks
- 2. Creare un nuovo Notebook con kernel Spark Scala
- 3. Inserire il seguente codice nelle celle:

```
%AddDeps org.apache.spark spark-sql-kafka-0-10_2.12 3.4.0 --
transitive
```

```
// Subscribe to test001 topic
val df = spark
.readStream
.format("kafka")
.option("kafka.bootstrap.servers", "kafka:9092")
.option("subscribe", "test001")
.load()
.selectExpr("CAST(key AS STRING)", "CAST(value AS STRING)")
.as[(String, String)]
```

```
import org.apache.spark.sql.streaming.Trigger
```

```
df.writeStream.format("console")
   .trigger(Trigger.ProcessingTime("5 seconds"))
   .outputMode("append")
   .start()
```

```
4. Mandare in esecuzione il notebook
```



5. Pubblicare un messaggio sul topic tramite web ui (vedere test precedenti)



#### 6 Data Governance

Il modulo di "*Data Governance*" è basato sul tool DataHub che permette di accedere alle funzionalità di data lineage e data exploration.

🖯 Manage						
						ح
Welcome back, <b>DataHub</b> .			Lat Analytics	<sub>ර</sub> ් Ingestion	🗒 Govern 🗸	\$ 0.
	Q Search Datasets, People, & more					
	Try searching for Explor	e all >				
	Apache Toree SampleSparkLobS dataset					
Explor	re your data					
	Datasets Pipelines					
	6 0					

## 6.1 Data Lineage

Per accedere alla funzionalità di data lineage, è possibile cliccare sia sulla voce Pipelines sia su una delle sorgenti dati che il tool visualizza.

La funzionalità permette di visualizzare il flusso logico di trasformazione all'interno della piattaforma.

Q Search Datasets	s, People, & more	
Try searching for		Explore all >
urn:li:container:2144	2a269a6b85d631c566f app-20230	921102726-0005 dataset
Explore your data	مچ Pipelines 2	
Platforms		



Per esempio, entrando sulla sezione "Explorer your data" -> Pipelines è possibile arrivare al dettaglio della stessa che può essere di più tipologie (nel nostro screen per esempio è di tipo Spark).

Ď	Q Search Datasets, People, & more		Select a View	Lu Analytics	് <sup>0</sup> Ingestion	🗒 Govern 🗸	暾	0
	Pipelines / spark / prod							
	prod							
	< Data Pipeline 🛠 Spark Apache Toree							
	< Data Pipeline 🙀 Spark SampleSparkJob\$							
		< 1 >						

Una volta visualizzato il dettaglio del job spark, cliccando sulla label Tasks verranno visualizzati tutti i task relativi al job eseguito.

Una volta entrati all'interno del task sarà possibile visualizzare la lineage specifica, entrando sulla sezione apposita tramite il pulsante "Visualize Lineage".

L Search Datasets, People, & more			Select a View $\vee$	Lat Analytics	ල් Ingestion [=	রু Govern ⊻ র্ষ্
ark > prod				(i) Details	ol <mark>o</mark> Lineage	2 upstream, 1 d
now Full Titles 🚺 Compress Lineage 🔊 🊺 Sh	now Columns					🛱 All'
<pre>prove room provide the statement of the statement of</pre>	with a summary of the summary of th	Minis (sour Jacobia) Salations grapping r_country Minis A C Min_1745	f A			

Una volta visualizzato il grafico di lineage è possibile abilitare la visualizzazione delle colonne per rendere più completa la trasformazione effettuata.

Inoltre, cliccando su una sorgente dati è possibile visualizzare la lineage a partire da quella sorgente come da screenshot successivo:



Search Datasets, People, & more		Select a View V Ltd Analytics 5
od > s3		(i) Details
Full Titles Compress Lineage () Show Columns		
	→ 🖓 🔤 🔤 🔤	
	ANISI   Interest E	
	Hide 🔨 Q 🔶 😽 🏧 🔤 🔤	
	Country	
	Birth rate(births/1800 pc	
	Current account balance	
	Death rate(deaths/1000 po	
	Electricity - consumption	
	Electricity - production(	
	Exports	
	GDP	
	< 1 2 3 4 5 >	

In questo caso si sta visualizzando che la sorgente dati (csv) su s3 è utilizzata da due applicazioni differenti.

#### *6.2* Data Lineage – Implementazione spark

All'interno della piattaforma, la Data Lineage su spark viene effettuata mediante una libreria scala/spark.

Il seguente esempio mostra come utilizzare la libreria quando si hanno più fonti dati, una trasformazione e un file di output da scrivere sul data lake:

• Lettura del primo dataset

```
val csv = "s3a://dataset/countries.csv"
val countries = spark.read.format("csv").option("delimiter",
";").option("header", "true").option("inferSchema",
"true").load(csv)
```

• Push del dataset come sorgente su datahub



```
import com.leonardo.spark.datahub.DataHubUtils
```

```
DataHubUtils.saveEntity(countries, csv)(spark)
```

• Lettura del secondo dataset

```
val csv_extra_ue = "s3a://dataset/countries_extra_ue.csv"
val countries_extra_ue =
spark.read.format("csv").option("delimiter", ";").option("header",
"true").option("inferSchema", "true").load(csv_extra_ue)
```

• Push del dataset come sorgente su datahub

DataHubUtils.saveEntity(countries\_extra\_ue, csv\_extra\_ue)(spark)

• Trasformazione e scrittura dell'output

```
countries.createOrReplaceTempView("COUNTRIES")
val countries_out = spark.sql("select sum(Area) as sum_area from
COUNTRIES group by Country")
val csv output = "s3a://dataset/csv/countries groupBy country"
```

countries\_out.write.mode("overwrite").format("csv").save(csv\_output)

• Push del dataset come output su datahub

DataHubUtils.saveEntity(countries\_out, csv\_output)(spark)

• Link del task spark con le sorgenti e il file di output

```
DataHubUtils.link(Seq(csv, csv_extra_ue),
spark.sparkContext.appName, spark.sparkContext.applicationId,
csv_output)(spark)
```

#### 6.3 Data Exploration

Il tool datahub deve essere abile a scansionare il datalake al fine di effettuare la collezione dei metadati su tutte le sorgenti usate.



Per far questo occorre andare sulla sezione Ingestion (in alto a destra) e creare una nuova sorgente mediante l'utilizzo del pulsante "Create new source".

A questo punto il tool guiderà l'utente sia nel caso in cui si debba importare i metadati da sorgenti già definite (via ui) sia nel caso di altre sorgenti supportate (da file di configurazione).

Q. Search Datasets, People, & more				Select a View	🗸 🔤 🖉 Ingestion 🖉 Govern
Ingestion	New Ingestion Source			×	
tule, and run DataHub ingestion sources.	1 Choose Type	2 Configure Reci	pe Schedule ing	gestion – (4) Finish up	
Secrets	Q. Search indestion sources				
new source 📿 Refresh	, conciniĝentances			_	All V Q Search sou
🔶 Name 🔶 Schedule			**	82	\$
S3 None	BigQuery	Redshift	Snowflake	Kafka	EDIT
	8	8	<b>*</b>	4	
	Looker	LookML	Tableau	PowerBl	
	dbt Cloud	MysqL	Postgres	Hive	

Nel caso in cui si vuole eseguire la scansione dei file all'interno del nostro datalake, ciò sarà possibile dentro la sezione "Other" usando un file di questo tipo:

In questo esempio datahub effettuerà la scansione di tutti i file supportati che risiedono dentro il bucket "dataset".

Tutti i file poi saranno ricercabili mediante la barra di ricerca semantica.



Di seguito un esempio di ricerca di un termine che il tool ha identificato come colonna di un file CSV.

U manaye							
Q E	lectricity	Select a View V	🔟 Analytics	<sub>ø</sub> ⊄ ingestion	🗒 Govern 🗸	ø	0-
Filter	Advanced	Showing 1 - 4 of 4 results					:
Type Datasets (4) Platform Diatorm Tom Tom Tom Tom Tom Diator Dia	^ ) ~	Detaset WASSS S3e.//dataset/countries.csv Matters column Bectricity - production(WMN) Dataset MASSS S3e.//dataset/countries_extra_ue.csv Sat.//dataset/countries_extra_ue.csv Matches column Bectricity - production(WMN) Dataset MASSS > bi dataset > bi un il contaner.21442a263a6085d531c566f3c6146c7a Countries.csv O Updated K wests app Matches column Bectricity - production(WMN) Dataset MASSS > bi dataset > bi un il contaner.21442a263a6085d531c566f3c6146c7a Countries.csv O Updated K wests app Matches column Bectricity - production(WMN) Dataset MASSS > bi dataset > bi un il contaner.21442a263a6085d531c566f3c6146c7a Countries_extra_ue.csv O Updated K wests app Matches column Bectricity - production(WMN) Dataset MASSS > bi dataset > bi un il contaner.21442a263a6085d531c566f3c6146c7a Countries_extra_ue.csv O Updated K wests app Matches column Bectricity - production(WMN) Dataset MASSS > bi dataset > bi un il contaner.21442a263a6085d531c566f3c6146c7a Countries_extra_ue.csv O Updated K wests app Matches column Bectricity - production(WMN) Dataset MASSS > bi dataset > bi un il contaner.21442a263a6085d531c566f3c6146c7a Countries_extra_ue.csv O Updated K wests app Dataset MASSS > bi dataset > bi un il contaner.21442a263a6085d531c566f3c6146c7a					
		Matches column Electricity - production (kNh)					