UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

Dipartimento di Scienze Fisiche, Informatiche e Matematiche

Corso di Laurea in Informatica

Progettazione e sviluppo di un'applicazione Big Data per l'analisi e l'elaborazione di tweet in real-time

RELATORE
Chiar.mo Professore
Riccardo Martoglia

CANDIDATO
Alessandro Pillo
MATR. 111759

Anno Accademico
2018/2019

Ambito di ricerca:
Big Data

Fonte: IBM

Ogni giorno creiamo oltre

2.500.000.000.000.000.000

(2.5 quintilioni) di byte.

Ogni minuto:

- ~ 500.000 nuovi tweet;
- ~ 49.380 nuove foto su Instagram;
- ~ 750.000 canzoni ascoltate su Spotify;
- ~ 176.220 chiamate su Skype;
- ~ 2.083.333 nuove snap su Snapchat.

Fonte: Domo

Paper

“Towards Tweet Content Suggestions For Museum Media Managers”

Professori:

- Riccardo Martoglia
- Federica Mandreoli
- Manuela Montangero
- Marco Furini

Progetto

- Esigenze e requisiti
- Studio preliminare
- Scelte progettuali
- Scelte implementative
- Valutazione

Esigenze e requisiti

- Progettazione di un'applicazione in Python
  - in grado di elaborare stream di tweet in real-time

- Costruzione di un modello predittivo in grado di:
  - evolvere nel tempo
  - effettuare previsioni riguardanti la popolarità o meno di un tweet

- Tweet riguardanti tematiche concernenti il mondo dell'arte

Esigenze e requisiti

- Progettazione di un'applicazione in Python
  - in grado di elaborare stream di tweet in real-time

- Costruzione di un modello predittivo in grado di:
  - evolvere nel tempo
  - effettuare previsioni riguardanti la popolarità o meno di un tweet

- Tweet riguardanti tematiche concernenti il mondo dell'arte

Esigenze e requisiti

- Progettazione di un'applicazione in Python
  - in grado di elaborare stream di tweet in real-time

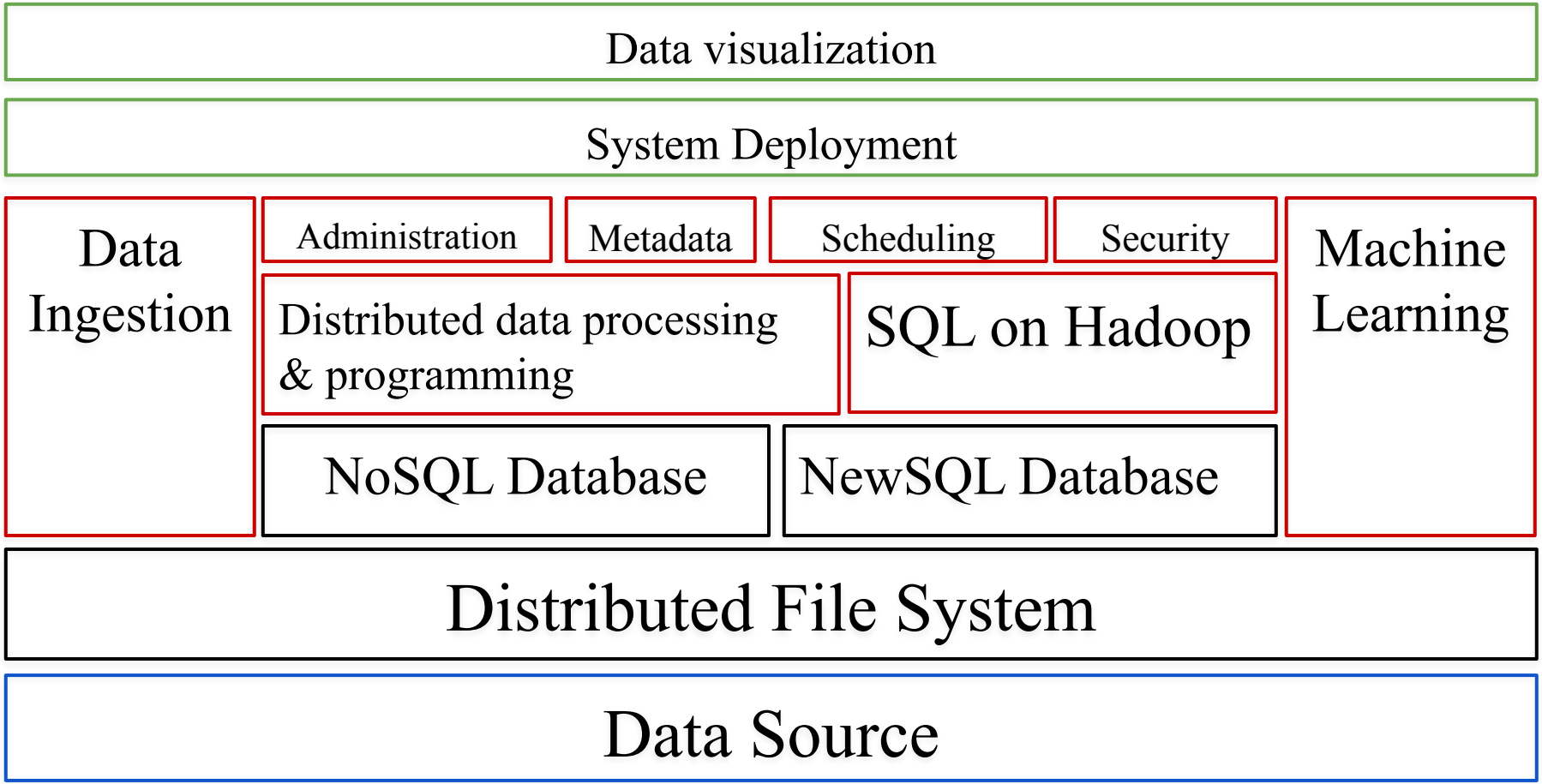
- Costruzione di un modello predittivo in grado di:
  - evolvere nel tempo
  - effettuare previsioni riguardanti la popolarità o meno di un tweet

- Tweet riguardanti tematiche concernenti il mondo dell'arte

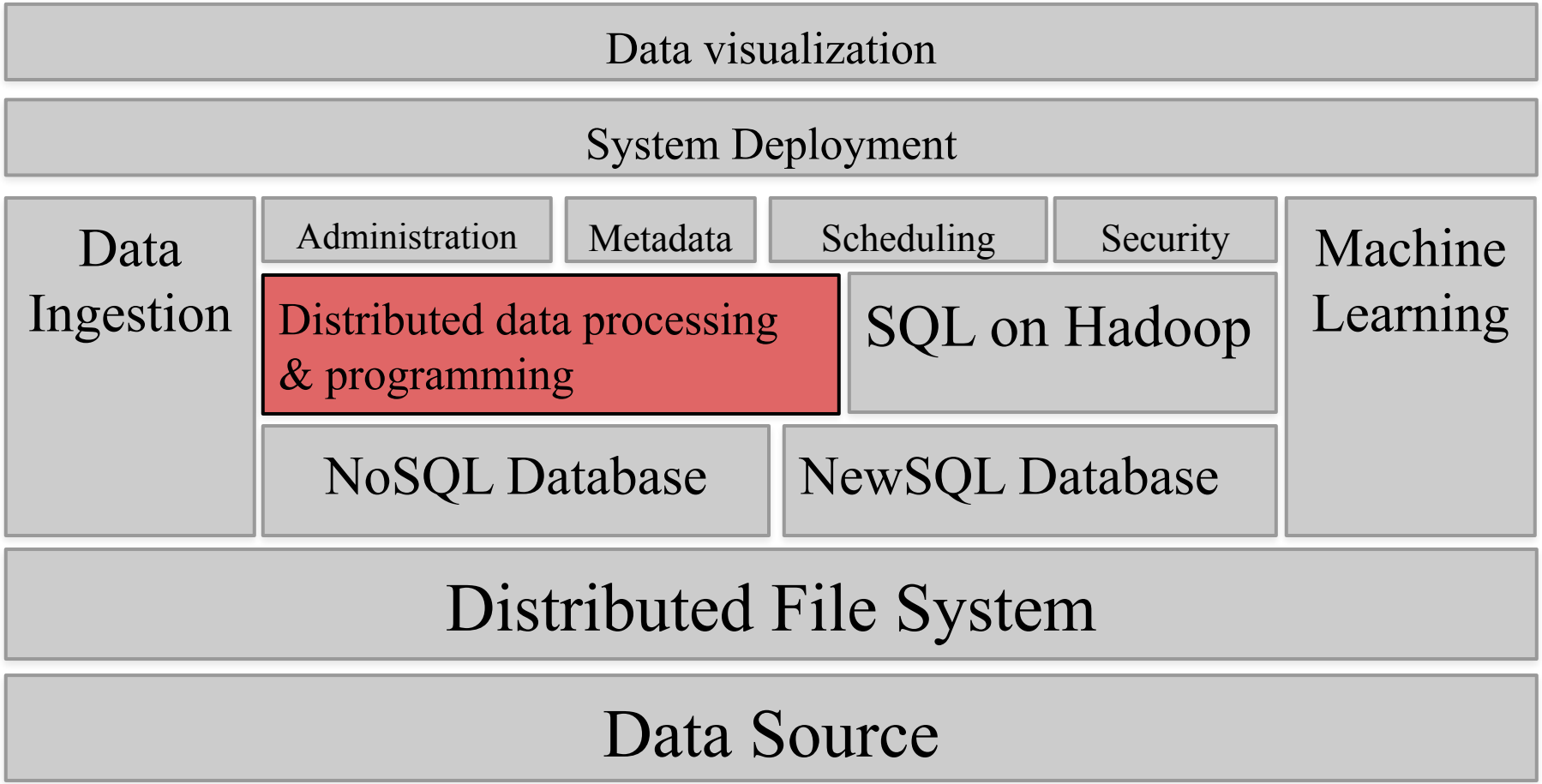
Progetto

- Esigenze e requisiti
- → Studio preliminare
- Scelte progettuali
- Scelte implementative
- Valutazione

Ecosistema Hadoop



Ecosistema Hadoop



Distributed data processing & programming

- Apache Apex
- Apache Beam
- Apache Hadoop
- Apache Flink
- Apache Samza
- Apache Spark
- Apache Storm
- Apache Tez
- Cloud Dataflow
- Google MillWheel
- IBM InfoSphere Streams
- Twitter Heron
- ....

Confronto

Parametri:

- Framework open source.
- Framework che implementa un engine di elaborazione.
- Possibilità di gestire il processing di dati stream.
- Integrazione nativa o estensibile con il machine learning su dati stream.
- Supporto al linguaggio di programmazione Python.
- Popolarità del framework.

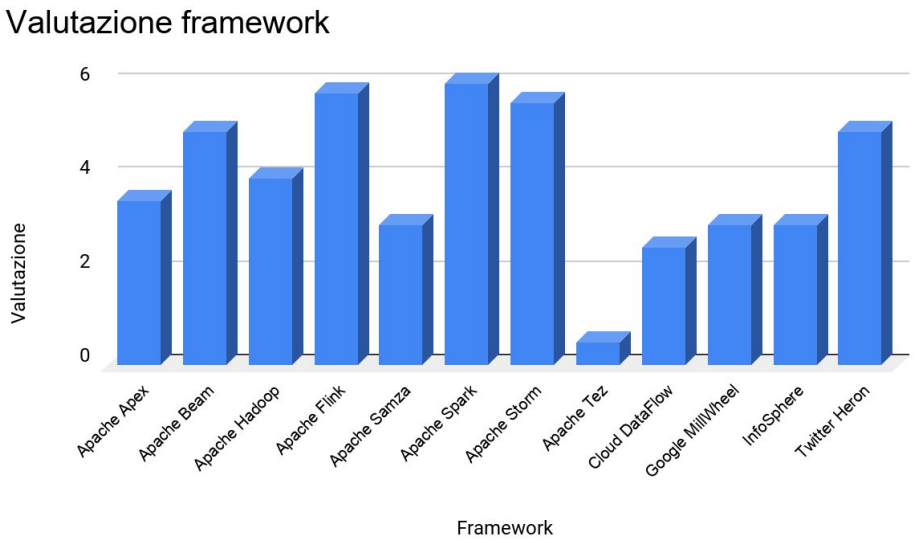


Grafico ottenuto assegnando ad ogni sistema, appartenente alla categoria “distributed data processing & programming”, un punteggio in base al fatto che esso soddisfi o meno ognuno dei parametri elencati.

Confronto

Parametri:

- Framework open source.
- Framework che implementa un engine di elaborazione.
- Possibilità di gestire il processing di dati stream.
- Integrazione nativa o estensibile con il machine learning su dati stream.
- Supporto al linguaggio di programmazione Python.
- Popolarità del framework.

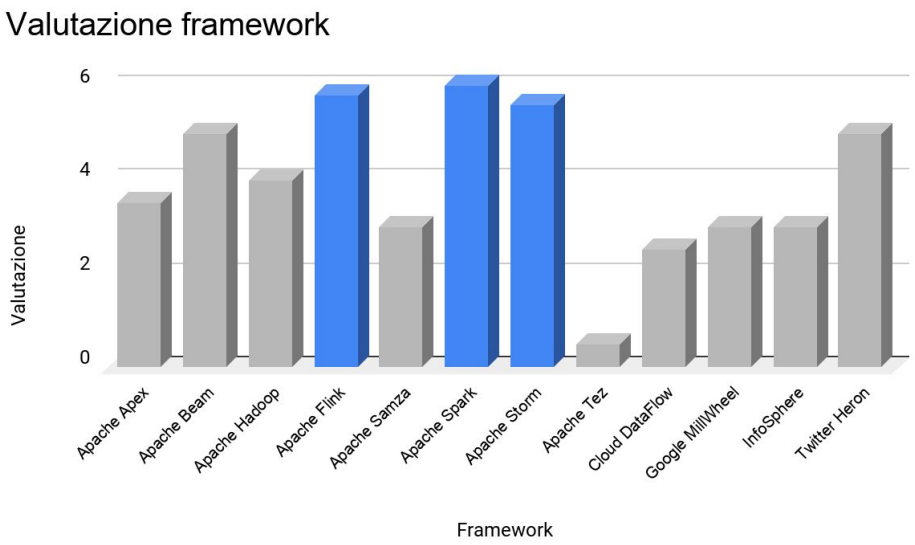


Grafico ottenuto assegnando ad ogni sistema, appartenente alla categoria “distributed data processing & programming”, un punteggio in base al fatto che esso soddisfi o meno ognuno dei parametri elencati.

Progetto

- Esigenze e requisiti
- Studio preliminare
- → Scelte progettuali
- Scelta dei parametri
- Valutazione

1 - Scelta dell’engine di elaborazione

Apache Flink

Supporta il Jython:

- documentazione ridotta;
- l’API DataStream è in versione beta e con bug noti;

Non integra una libreria per l’online ML.

1 - Scelta dell’engine di elaborazione

Apache Flink

Supporta il Jython:

- documentazione ridotta;
- l’API DataStream è in versione beta e con bug noti;

Non integra una libreria per l’online ML.

Apache Storm

E’ multilinguaggio, tuttavia richiede l’implementazione di opportune classi in Java per invocare i metodi scritti in altri linguaggi.

La Stream API è in versione sperimentale.

Non integra una libreria per l’online ML.

1 - Scelta dell’engine di elaborazione

Apache Flink

Supporta il Jython:

- documentazione ridotta;
- l’API DataStream è in versione beta e con bug noti;

Non integra una libreria per l’online ML.

VS.

Apache Spark

Le principali funzionalità che esso offre sono disponibili anche in Python.

Documentazione dettagliata.

La libreria MLlib supporta l’online ML.

VS.

Apache Storm

E’ multilinguaggio, tuttavia richiede l’implementazione di opportune classi in Java per invocare i metodi scritti in altri linguaggi.

La Stream API è in versione sperimentale.

Non integra una libreria per l’online ML.

2 - Scelta dell’algoritmo di ML

- Tipologia di ML:
  - online machine learning
- Libreria MLlib:
  - metodo *StreamingLogisticRegression()*

Pipeline concettuale

Sorgente

Apache Kafka

Data ingestion

Engine di elaborazione

MLlib

Online ML

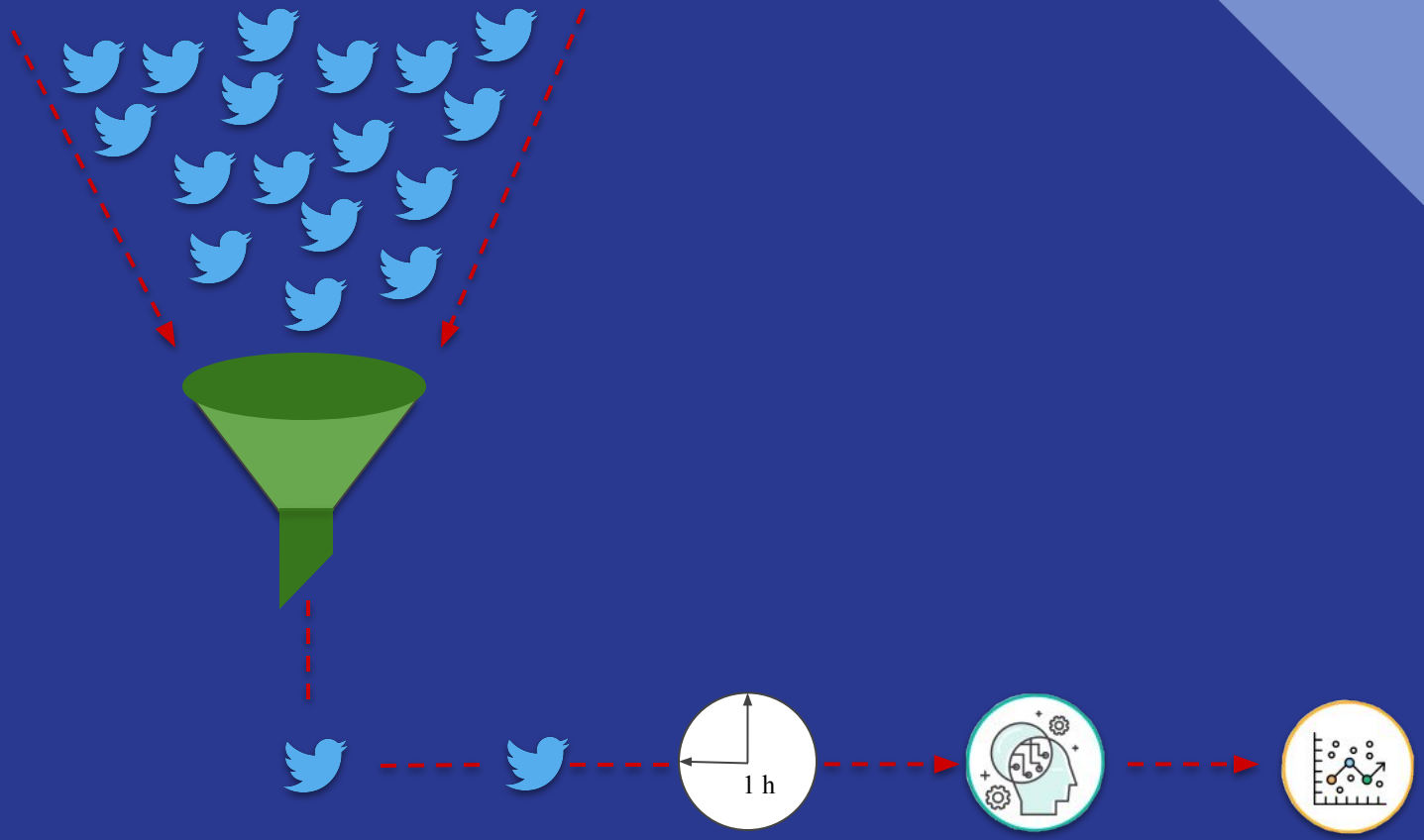
Progetto

- Esigenze e requisiti
- Studio preliminare
- Scelte progettuali
- → Scelte implementative
- Valutazione

Elaborazione dello stream

1. Filtraggio iniziale

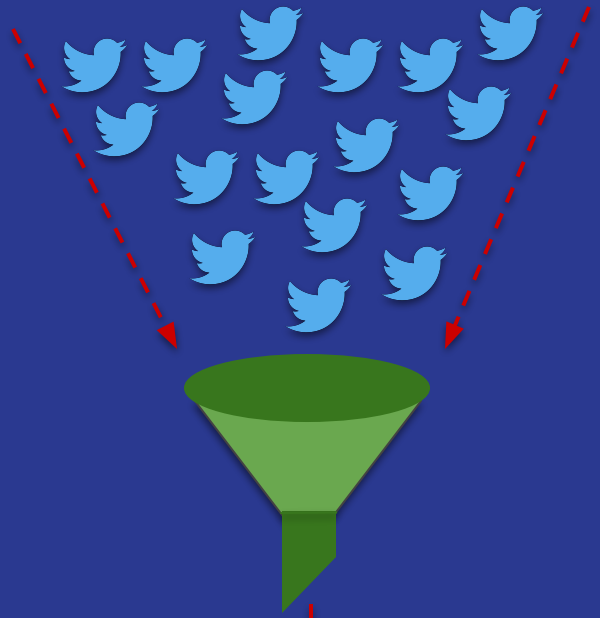
- *tematica inerente l'arte;*
- *tweet "originali";*
- *follower tra 1.000 e 30.000;*



Elaborazione dello stream

1. Filtraggio iniziale

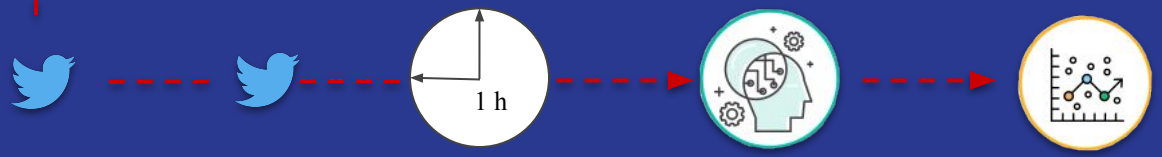
- *tematica inerente l'arte;*
- *tweet "originali";*
- *follower tra 1.000 e 30.000;*



2. Tempo di attesa

- *Valore: 1 ora*

Per valutare il numero di like ottenuto.



Elaborazione dello stream

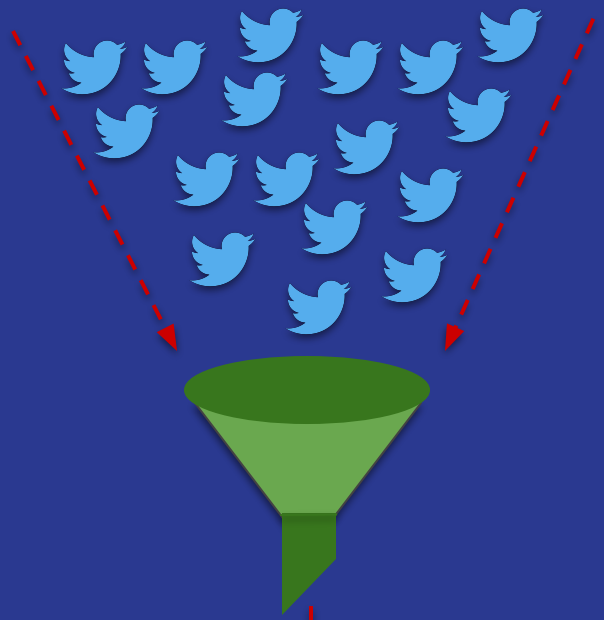
1. Filtraggio iniziale

- *tematica inerente l'arte;*
- *tweet "originali";*
- *follower tra 1.000 e 30.000;*

2. Tempo di attesa

- *Valore: 1 ora*

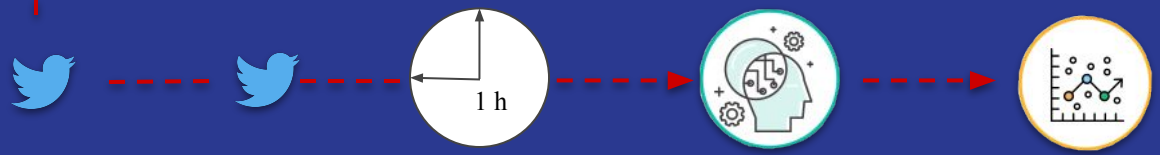
Per valutare il numero di like ottenuto.



3. Filtraggio dataset per ML

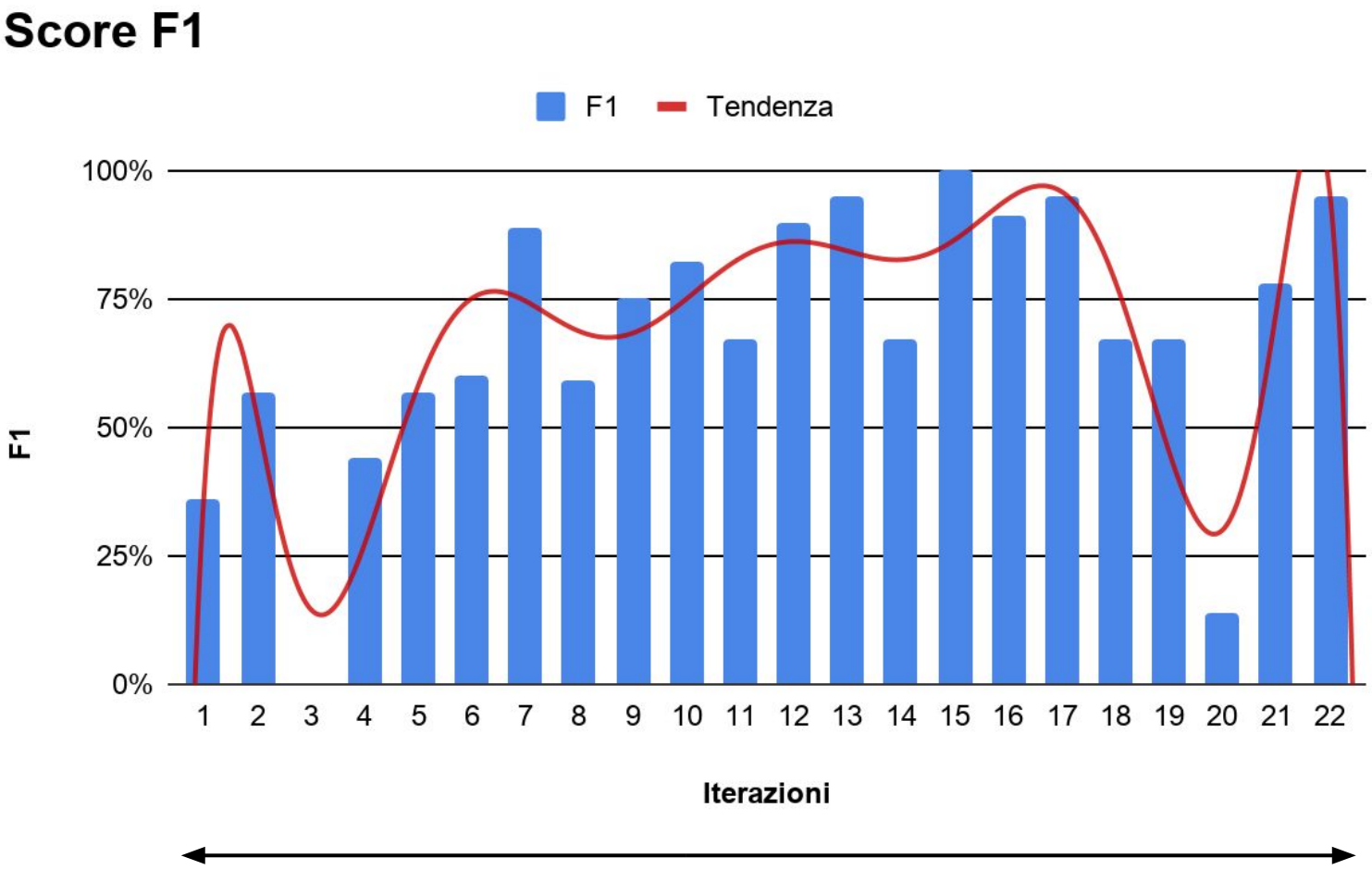
- *Soglie: like <= 3 e like >= 15*

Per addestrare il modello sulla porzione di dati maggiormente significativa.

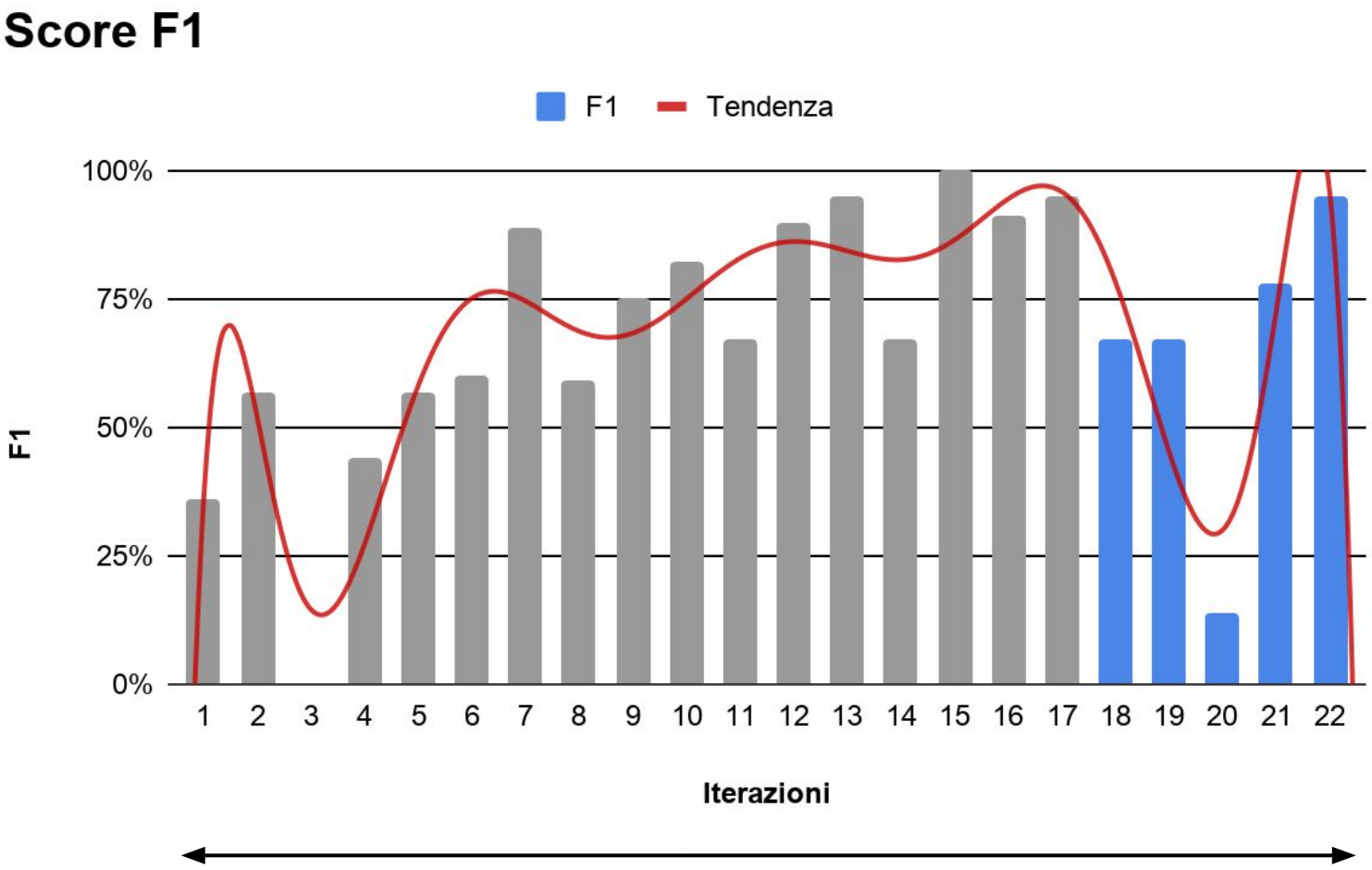


Progetto

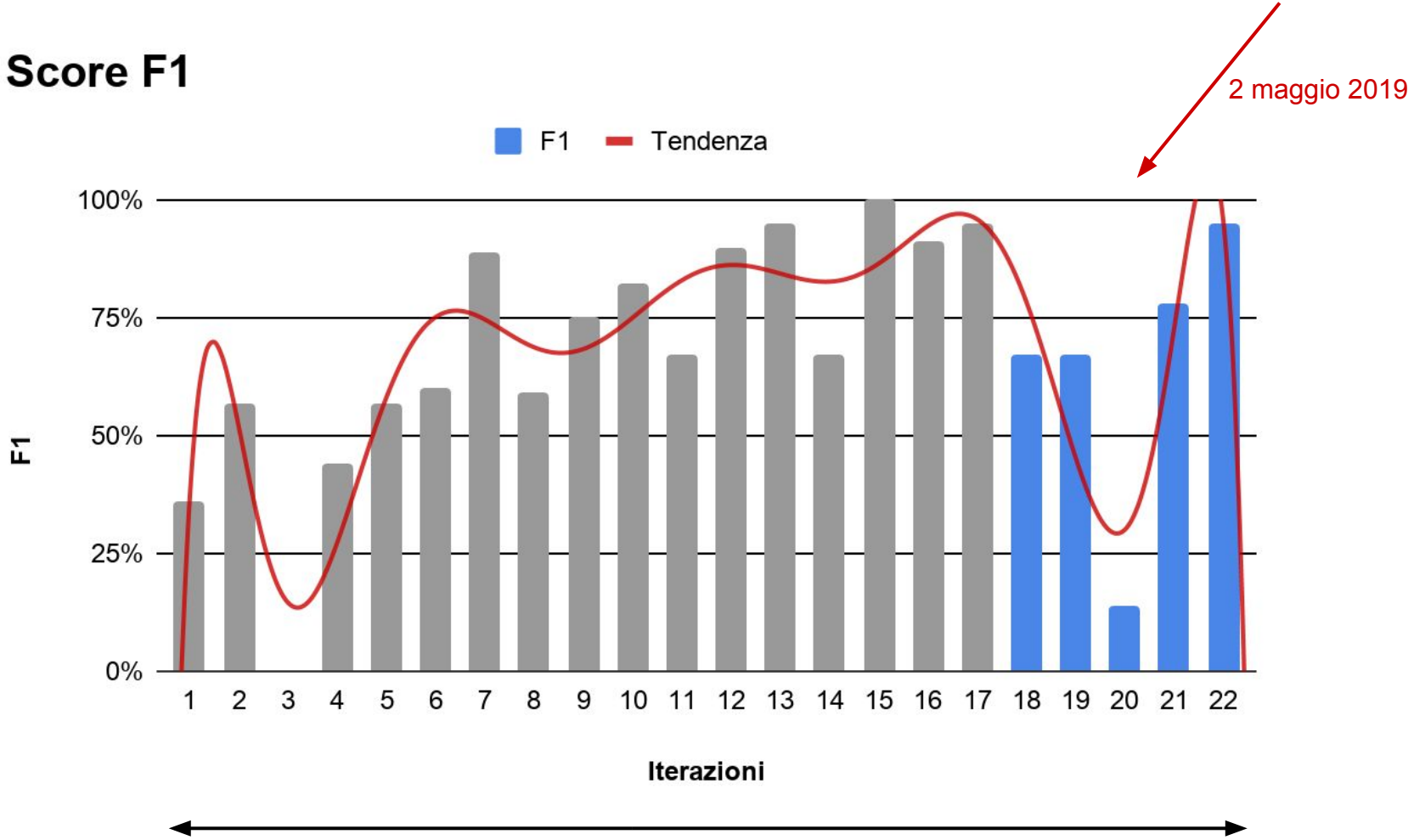
- Esigenze e requisiti
- Studio preliminare
- Scelte progettuali
- Scelte implementative
- Valutazione


Score F1
F1
Tendenza
100%
75%
50%
25%
0%
F1
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22
Iterazioni

1 settimana


Score F1
F1
Tendenza
100%
75%
50%
25%
0%
F1
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22
Iterazioni

1 settimana


Score F1
2 maggio 2019
F1
Tendenza
100%
75%
50%
25%
0%
F1
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22
Iterazioni

1 settimana

Input rate

Input Rate
Avg: 9397,30 records/sec
records/sec
20,000.00
15,000.00
10,000.00
5,000.00
0.00
20:48:37
21:05:16

Conclusioni

- E’ stata delineata la struttura dell’ecosistema di framework caratterizzanti il panorama dei Big Data.
- Sono stati approfonditi i principali framework utili al fine di analizzare, elaborare e interrogare dati.
- E’ stata progettata un’applicazione capace di elaborare uno stream di dati in real-time.
- E’ stato costruito un modello predittivo in grado di evolvere nel tempo e si è verificato sperimentalmente le sue capacità di adattamento e i vantaggi introdotti dall’online machine learning.

Sviluppi futuri

- Costruire un modello predittivo in grado di fare previsioni riguardanti il numero di like che un determinato tweet dovrebbe riuscire a raggiungere sulla piattaforma.

Grazie per l’attenzione