



UN FRAMEWORK PER IL NATURAL LANGUAGE PROCESSING: ANALISI PRESTAZIONALE PER LA RISOLUZIONE DI TASK DI TEXT CLASSIFICATION E NAMED ENTITY RECOGNITION IN AMBIENTE DISTRIBUITO

Tesi di Laurea Triennale

Relatore:
Prof. Roberto Basili

Candidato:
Manuel Di Lullo

Correlatore:
Prof. Danilo Croce

Outline

- Elaborazione di informazione non strutturata nell'epoca dei big data
- Obiettivo della tesi: studio e implementazione di un approccio basato su calcolo distribuito
- Soluzione proposta ed implementata
 - HDFS, MapReduce, YARN, Spark, **Spark NLP**
- Valutazione sperimentale
 - Text Classification e Named Entity Recognition
- Conclusioni

Un mondo fatto di dati



319.6 miliardi

Di mail inviate nel 2021

100 miliardi

Di messaggi Whatsapp nel 2021

1.8 miliardi

Di utenti Facebook nel 2021

463 Exabyte

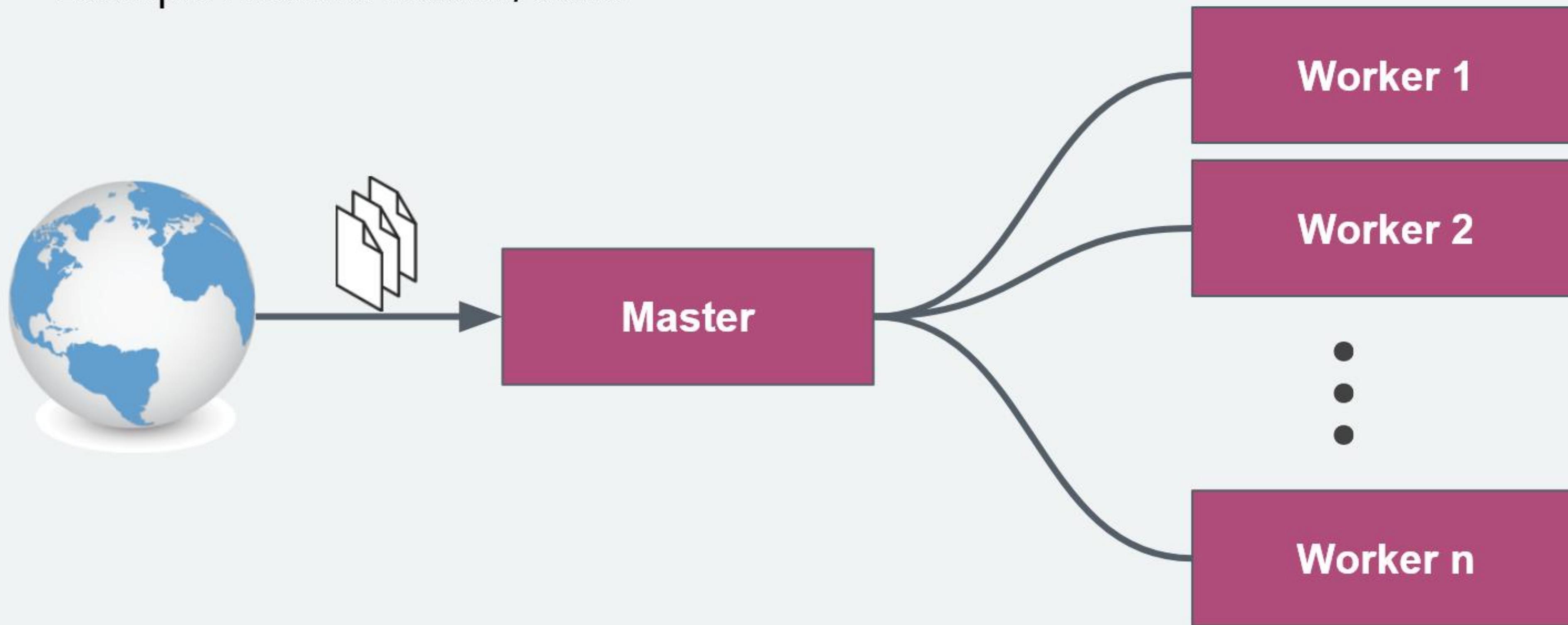
Di dati generati ogni giorno nel 2025

Fonte: statista.com

Il calcolo distribuito

Campo dell'informatica che studia i sistemi distribuiti, ovvero sistemi che consistono in numerosi computer che interagiscono tra loro attraverso una rete al fine di raggiungere un obiettivo comune.

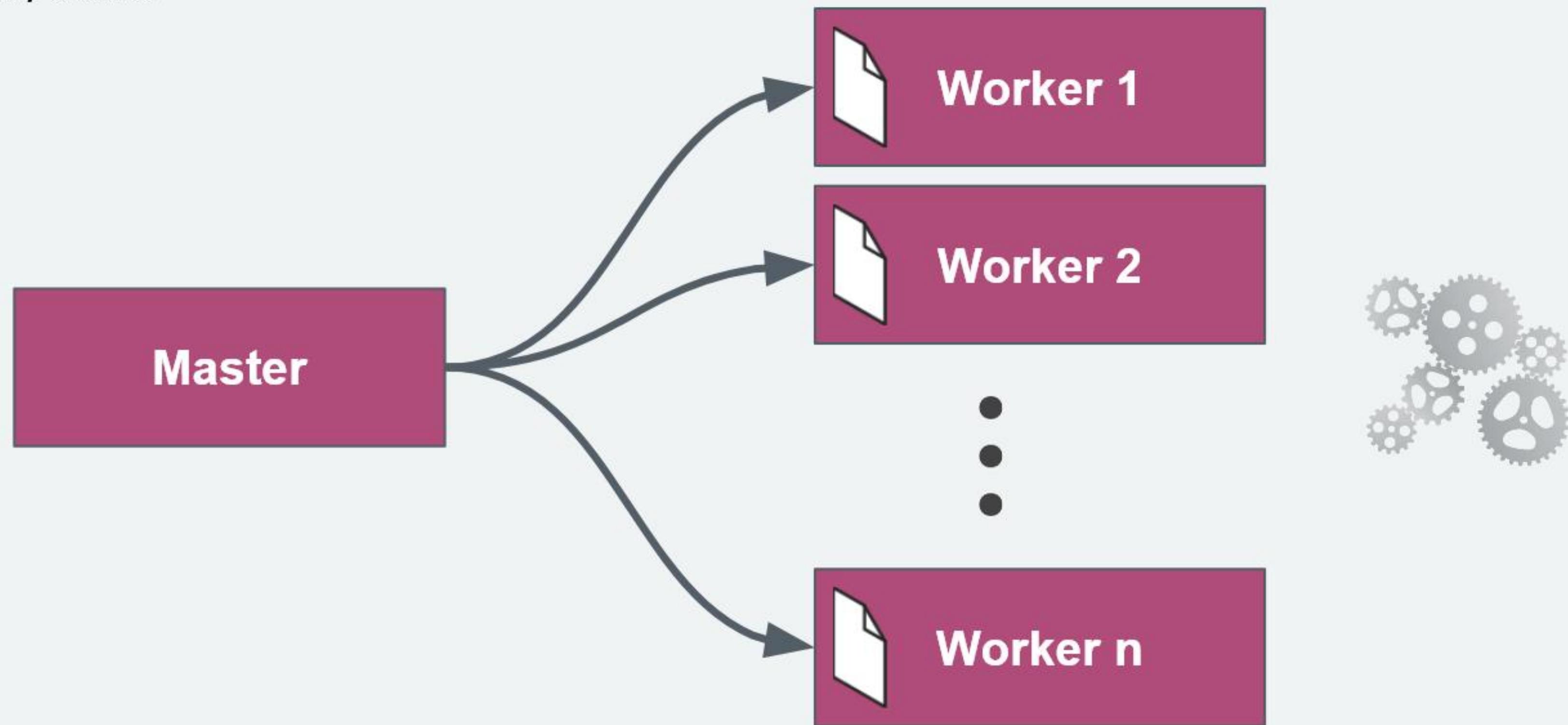
Esempio modello master/slave



Il calcolo distribuito

Campo dell'informatica che studia i sistemi distribuiti, ovvero sistemi che consistono in numerosi computer che interagiscono tra loro attraverso una rete al fine di raggiungere un obiettivo comune.

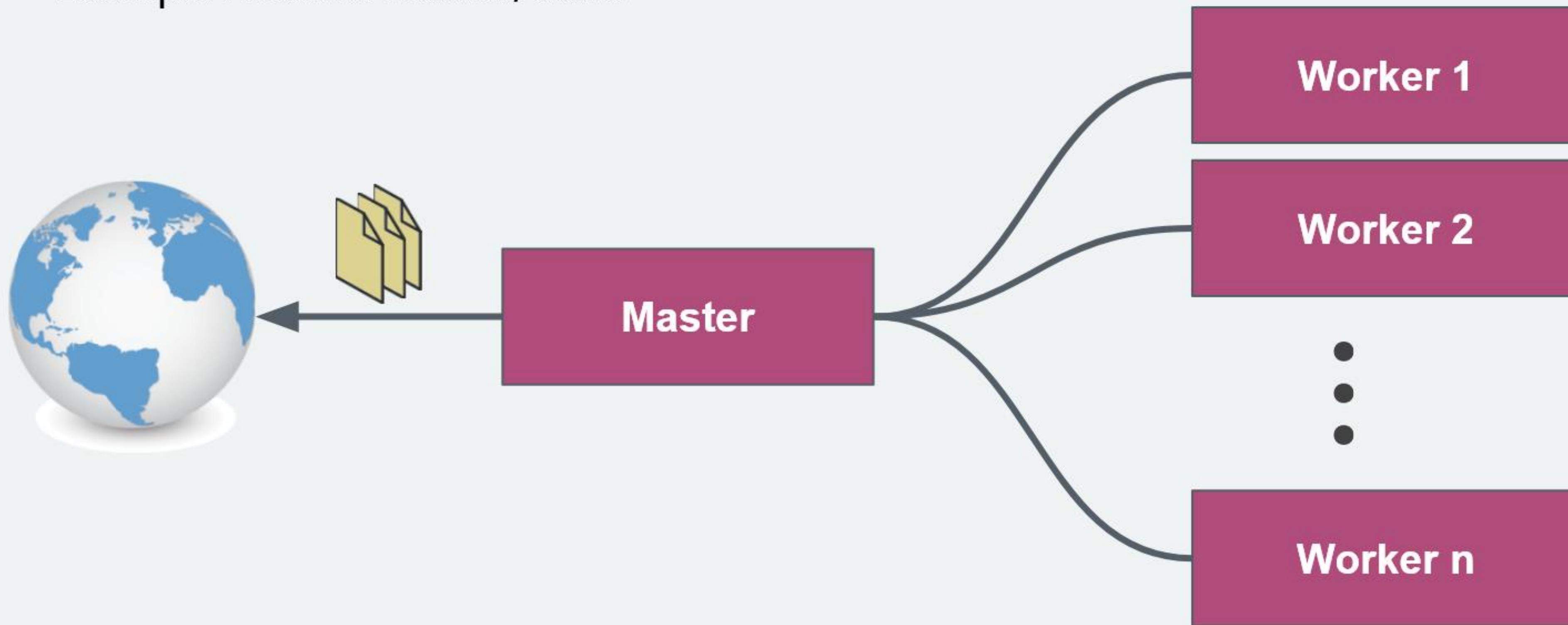
Esempio modello master/slave



Il calcolo distribuito

Campo dell'informatica che studia i sistemi distribuiti, ovvero sistemi che consistono in numerosi computer che interagiscono tra loro attraverso una rete al fine di raggiungere un obiettivo comune.

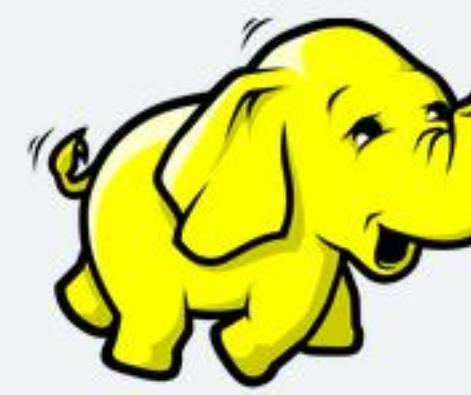
Esempio modello master/slave



Obiettivi della tesi

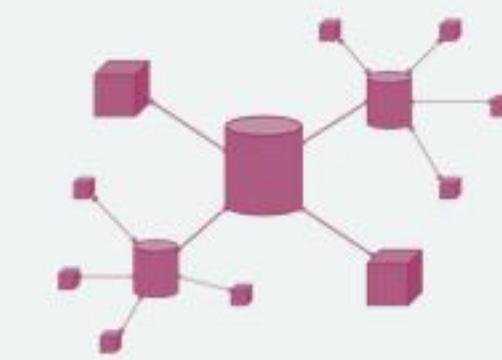
1

Studio e definizione
dell'ecosistema
Hadoop per il calcolo
distribuito di dati non
strutturati



2

Creazione di un
ambiente distribuito
con l'ausilio di
Hadoop e Apache
Spark



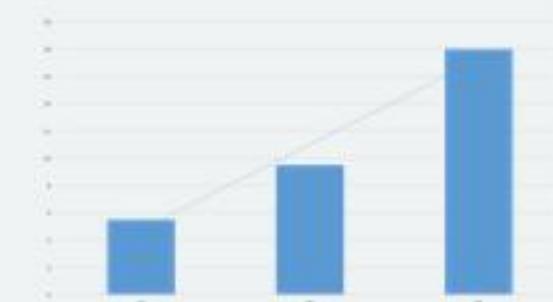
3

Studio di modelli per
la risoluzione di task
di natura linguistica
e del framework
Spark NLP



4

Valutazione
sperimentale dei
modelli. Misura
dell'accuratezza e
della scalabilità
della soluzione.



Un framework per il calcolo distribuito

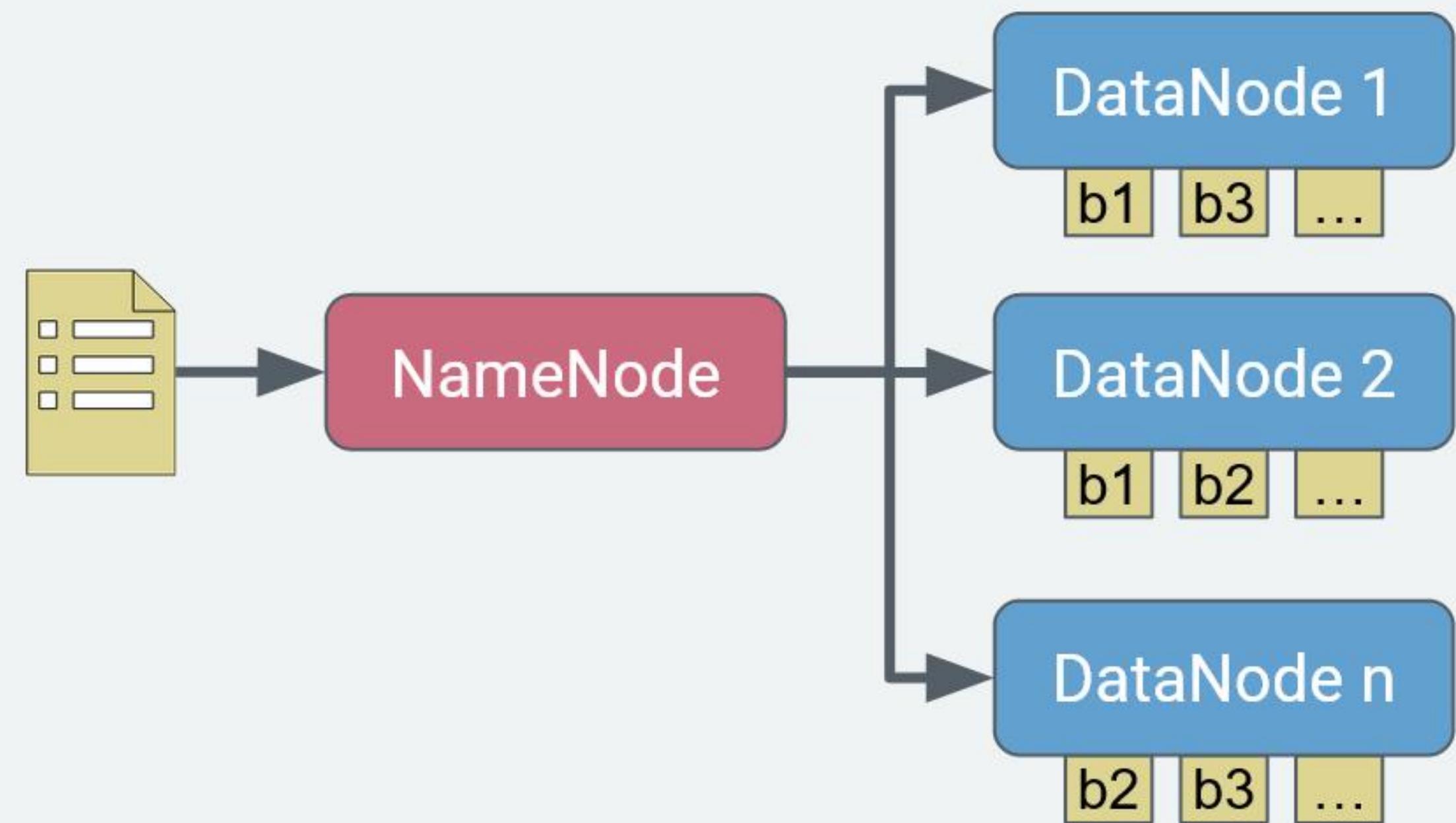
- Hadoop è un framework che consente l'elaborazione distribuita di grandi insiemi di dati in ambiente distribuito
- In questa tesi sono stati trattati i seguenti moduli:
 - **Hadoop Distributed File System (HDFS)**: per la gestione dei file in ambiente distribuito
 - **Hadoop MapReduce**: per l'elaborazione distribuita dei dati
 - **Hadoop Yet Another Resource Negotiator (YARN)**: per la gestione delle risorse



Hadoop Distributed File System

Gestione dei file in ambiente distribuito

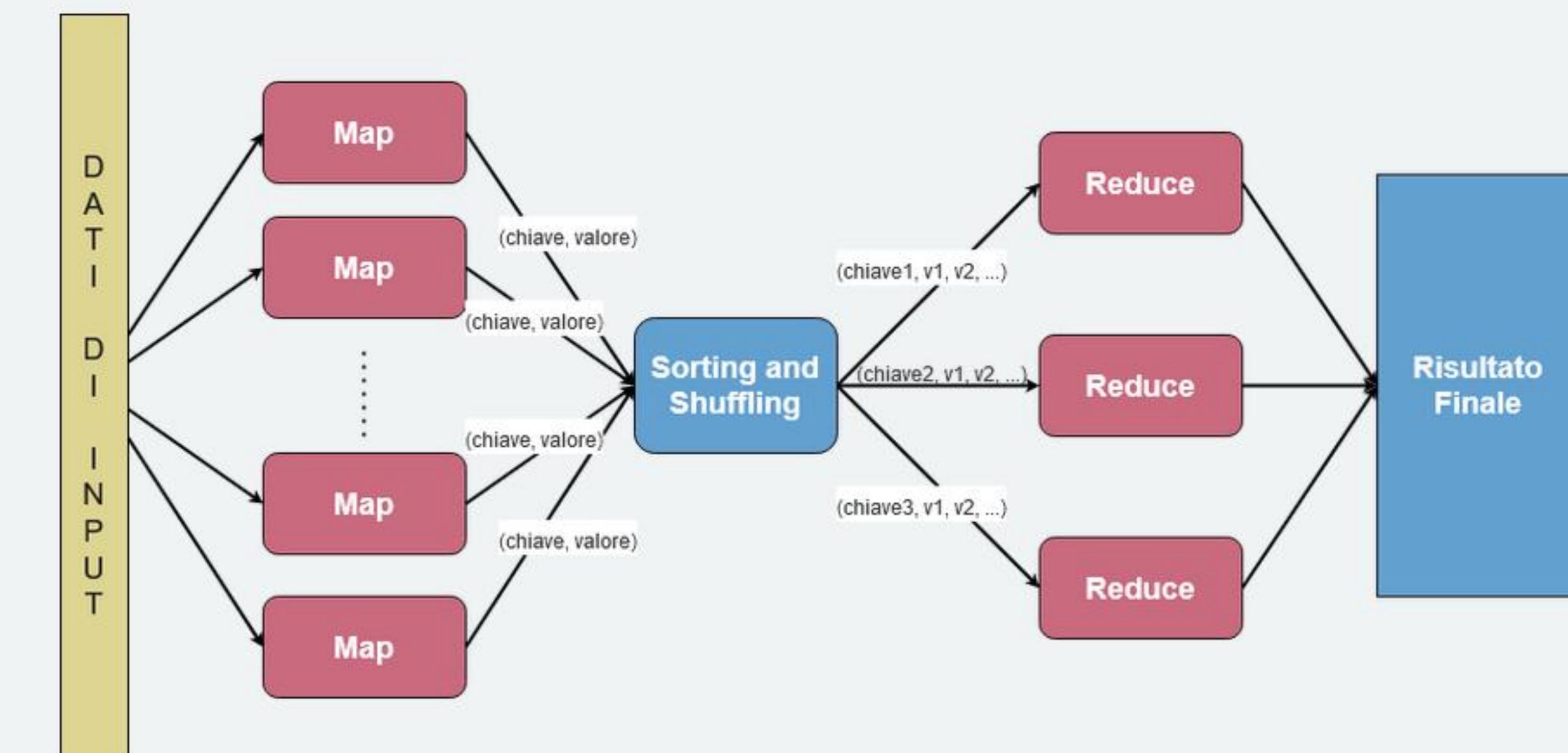
- File System gerarchico e distribuito
- Utilizzato per far sì che tutti i nodi del cluster abbiano una visione condivisa dei dati e dei modelli
- Altamente tollerante agli errori
 - File divisi in blocchi e spartiti tra le macchine
 - Replicazione dei blocchi



Hadoop MapReduce

Elaborazione di dati in parallelo

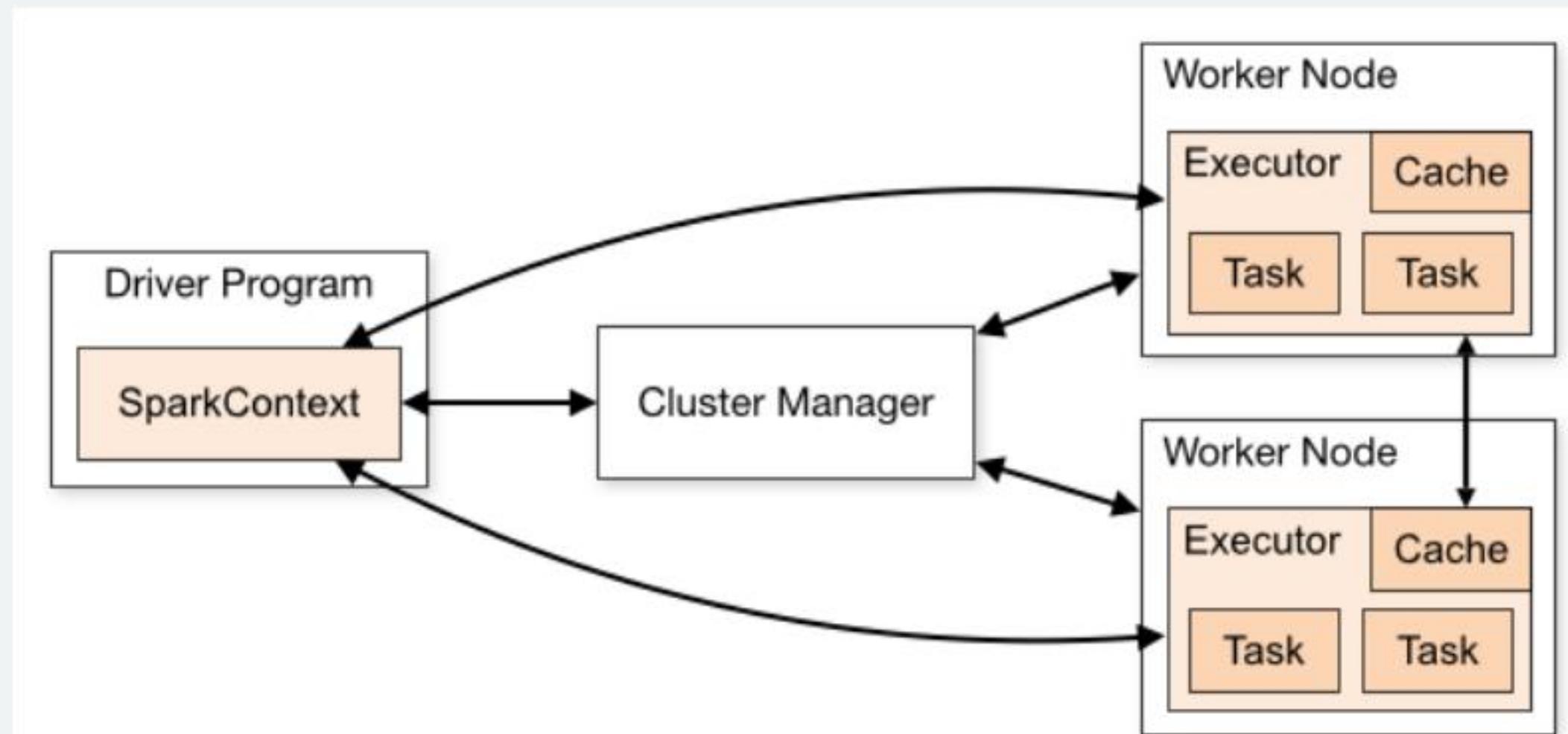
- MapReduce si occupa dell'elaborazione distribuita dei file memorizzati su HDFS.
- Composto principalmente da due funzioni: **Map e Reduce**
- *Divide et impera:*
 - Il nodo master divide l'operazione di calcolo in jobs e li assegna ai vari nodi del sistema
 - I jobs sono processati in modo autonomo sui nodi worker
 - Il master riduce i singoli risultati ottenuti dai nodi worker ad un unico risultato finale



Apache Spark



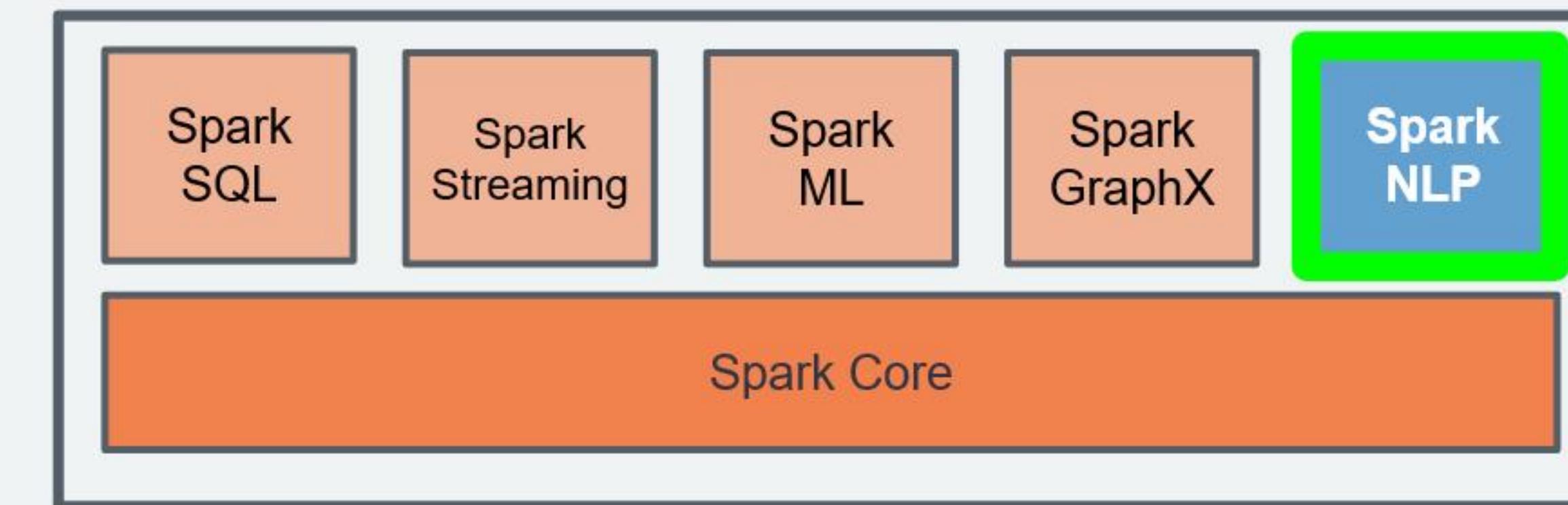
- Sistema di elaborazione distribuita open source.
- Utilizzato come alternativa a MapReduce
- Eredita l'architettura Master-Slave di MapReduce e risolve problemi presenti in quest'ultimo
- Compatibile con YARN e HDFS
 - YARN alloca e monitora le risorse necessarie all'applicazione
 - Spark sfrutta modelli e file memorizzati su HDFS



Apache Spark

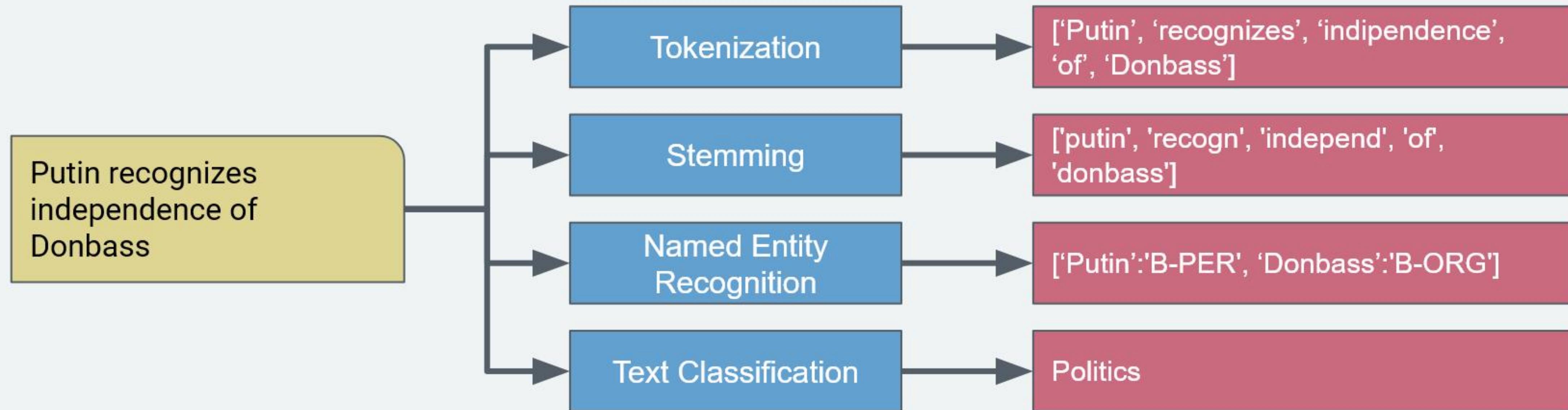


- Sistema di elaborazione distribuita open source.
- Utilizzato come alternativa a MapReduce
- Eredita l'architettura Master-Slave di MapReduce e risolve problemi presenti in quest'ultimo
- Compatibile con YARN e HDFS
 - YARN alloca e monitora le risorse necessarie all'applicazione
 - Spark sfrutta modelli e file memorizzati su HDFS
- Spark Core API e librerie per analisi dati



Spark NLP

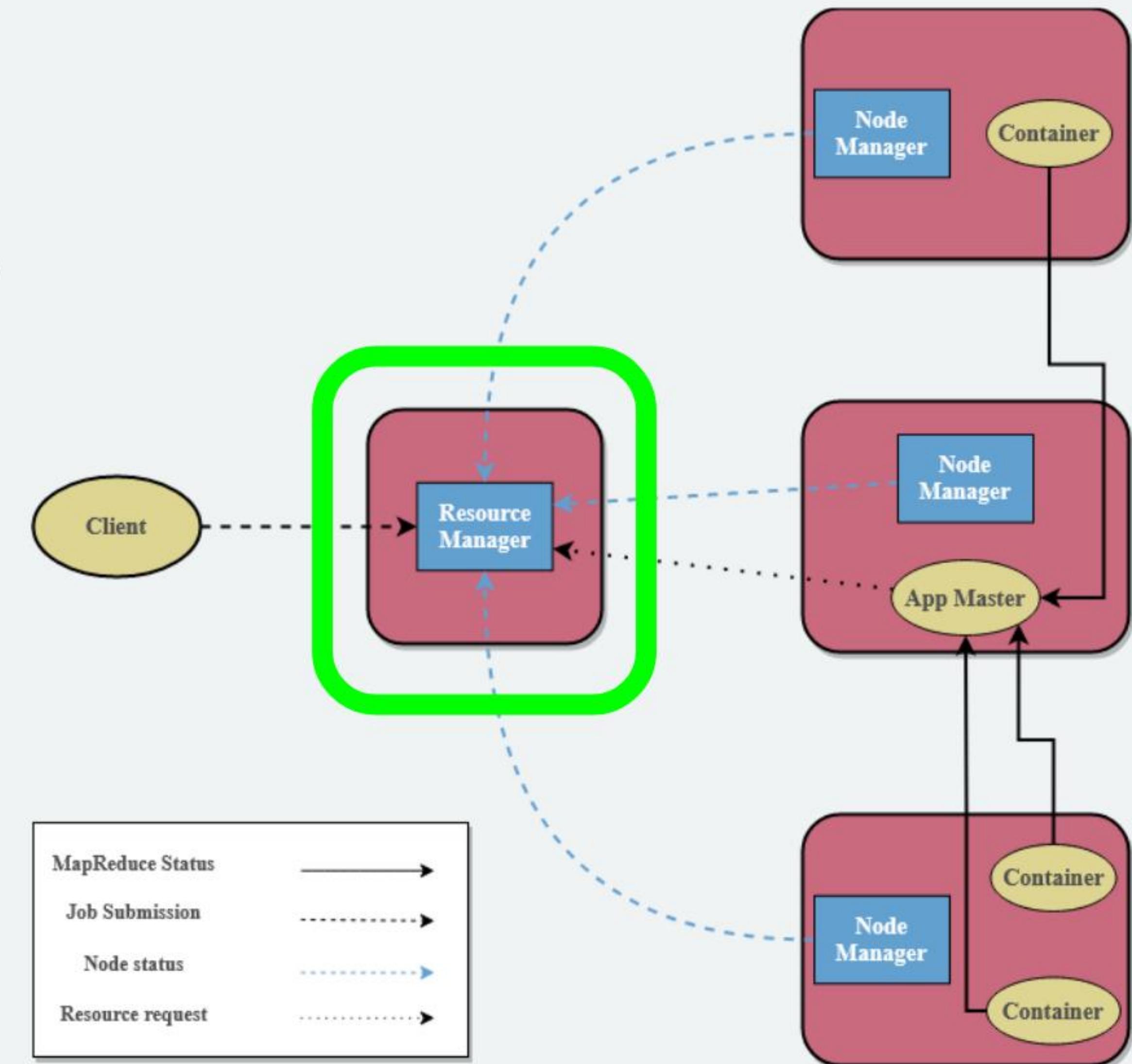
- Spark NLP è una libreria di elaborazione del linguaggio naturale open source.
- È costruita su Apache Spark e Spark ML ed è compatibile con framework per l'elaborazione distribuita come Hadoop.
- Implementa Machine Learning pipelines che rappresentano gli step che un modello deve compiere
- Dispone di decine di annotatori, molti dei quali sfruttano reti neurali per la classificazione di testi.



Hadoop YARN

Gestione delle risorse e scheduling dei jobs

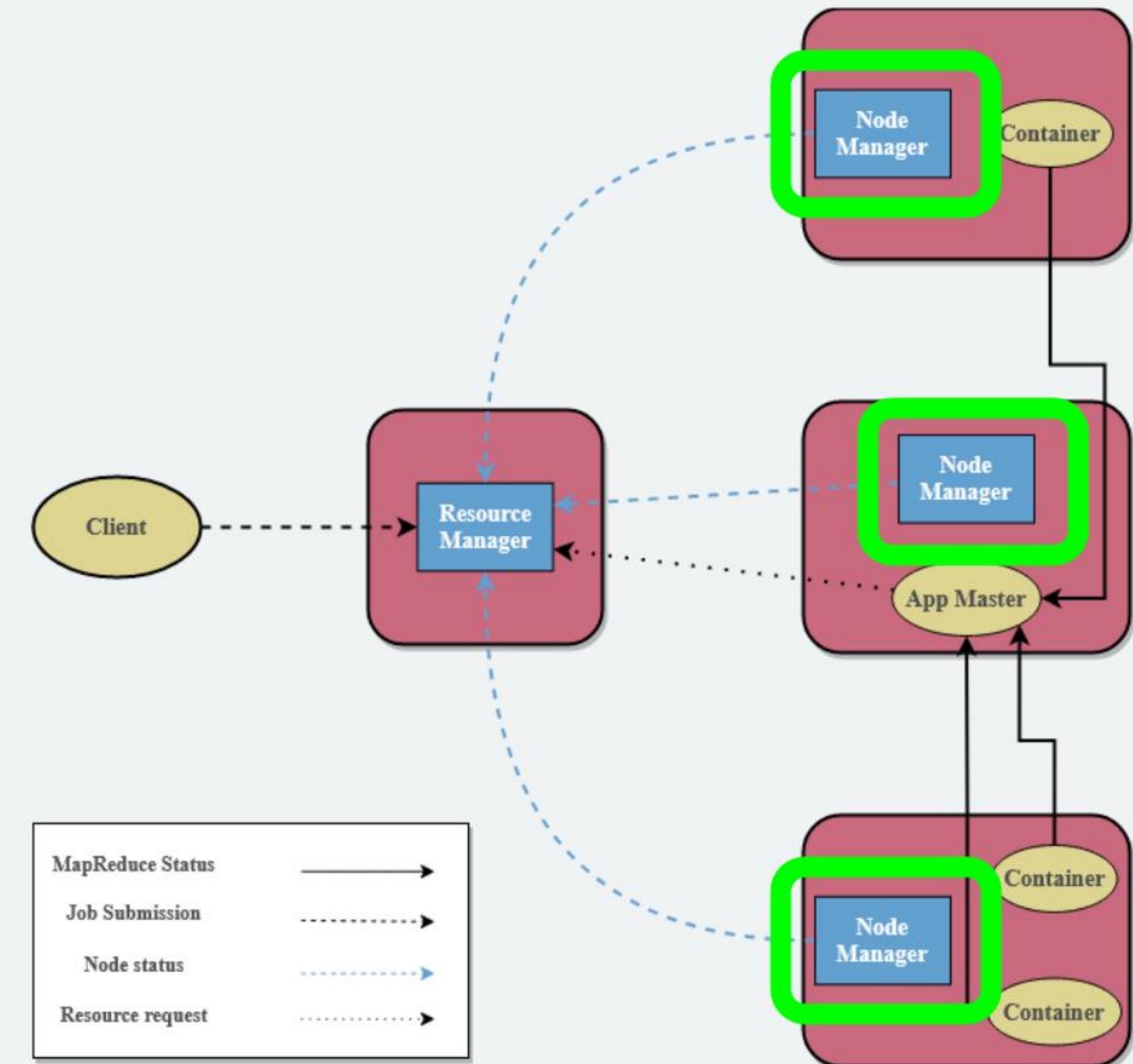
- Scheduling dei jobs e la gestione delle risorse del cluster.
 - Registra le applicazioni e le aggiunge alla coda dei job da eseguire
 - Assegna le risorse per ogni applicazione
 - Monitora lo stato delle risorse su ogni nodo
- Suddiviso in tre componenti:
 - **ResourceManager**: processo master di YARN



Hadoop YARN

Gestione delle risorse e scheduling dei jobs

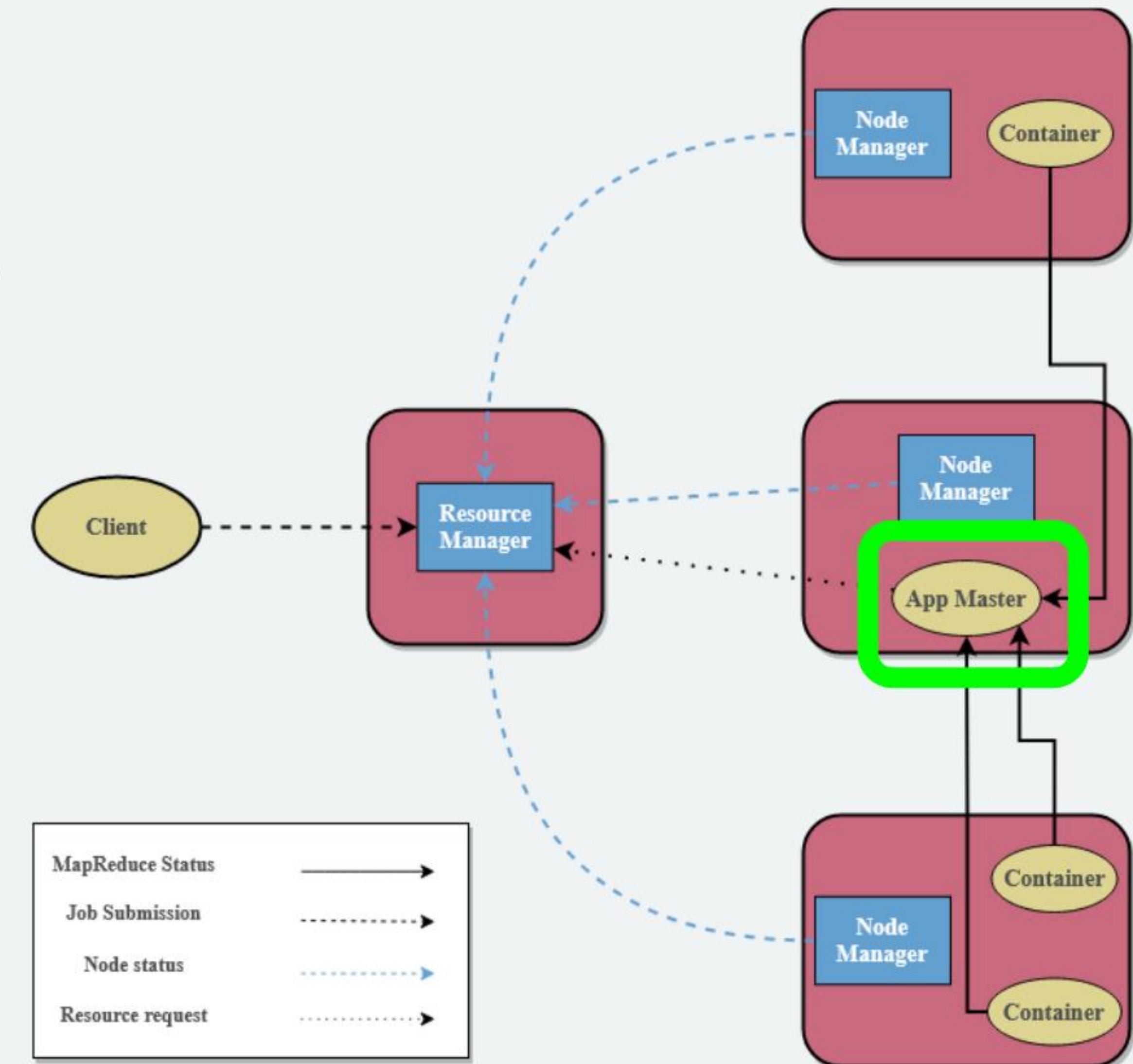
- Scheduling dei jobs e la gestione delle risorse del cluster.
 - Registra le applicazioni e le aggiunge alla coda dei job da eseguire
 - Assegna le risorse per ogni applicazione
 - Monitora lo stato delle risorse su ogni nodo
- Suddiviso in tre componenti:
 - **ResourceManager**: processo master di YARN
 - **NodeManager**: responsabile del monitoraggio delle risorse per ogni macchina.



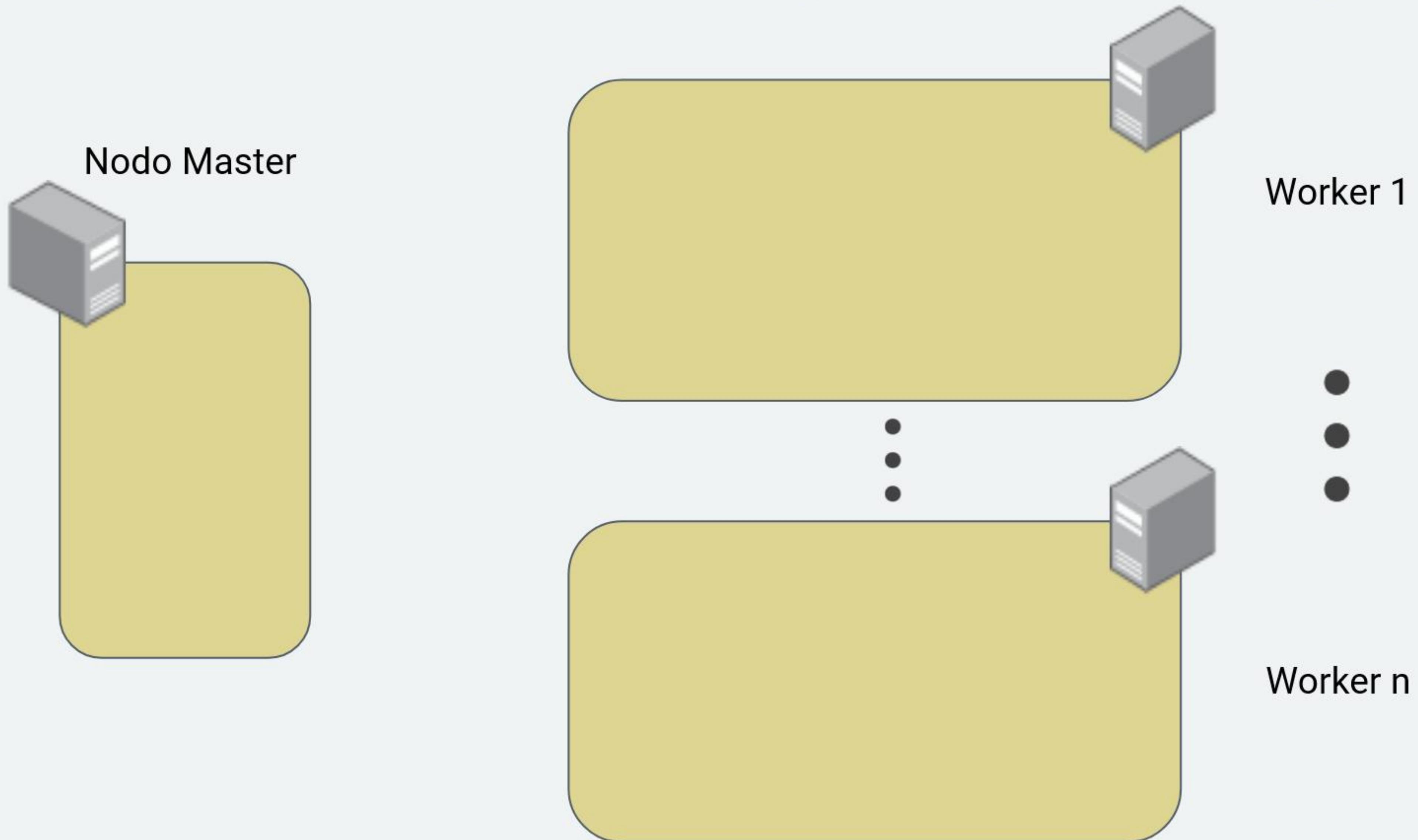
Hadoop YARN

Gestione delle risorse e scheduling dei jobs

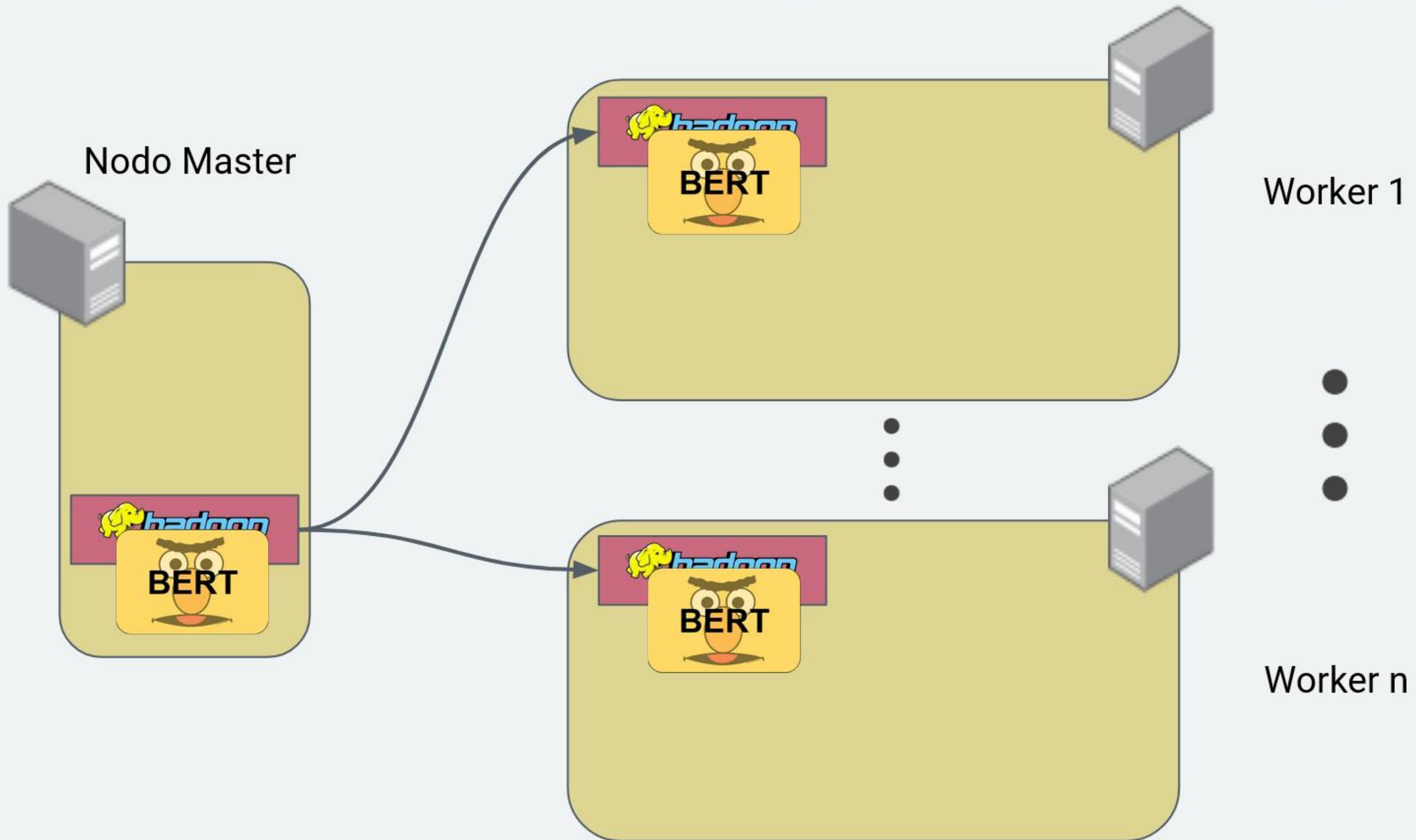
- Scheduling dei jobs e la gestione delle risorse del cluster.
 - Registra le applicazioni e le aggiunge alla coda dei job da eseguire
 - Assegna le risorse per ogni applicazione
 - Monitora lo stato delle risorse su ogni nodo
- Suddiviso in tre componenti:
 - **ResourceManager**: processo master di YARN
 - **NodeManager**: responsabile del monitoraggio delle risorse per ogni macchina.
 - **ApplicationMaster**: negozia le risorse dal ResourceManager e lavora con i NodeManager per eseguire e monitorare i job.



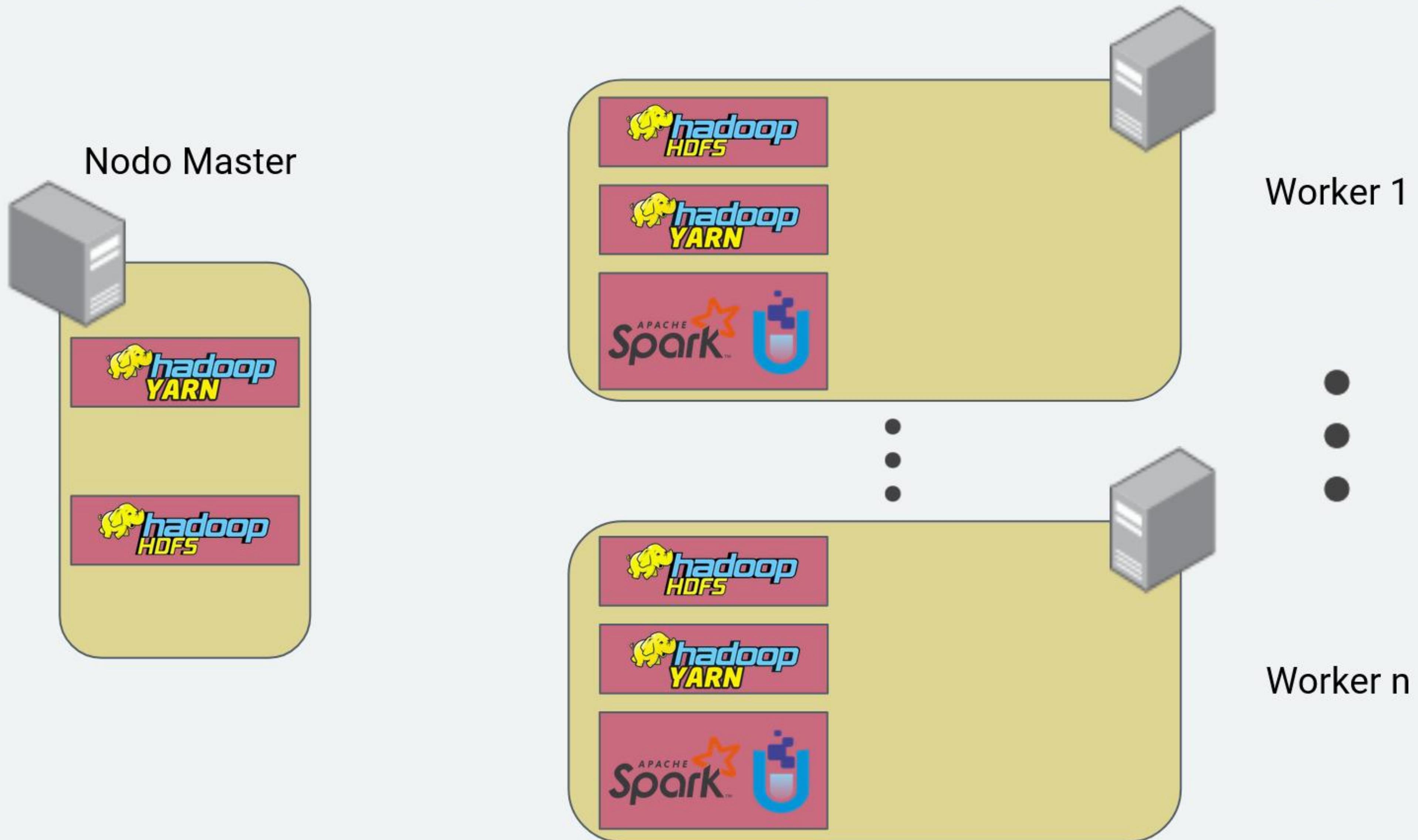
Calcolo distribuito con Spark e Hadoop



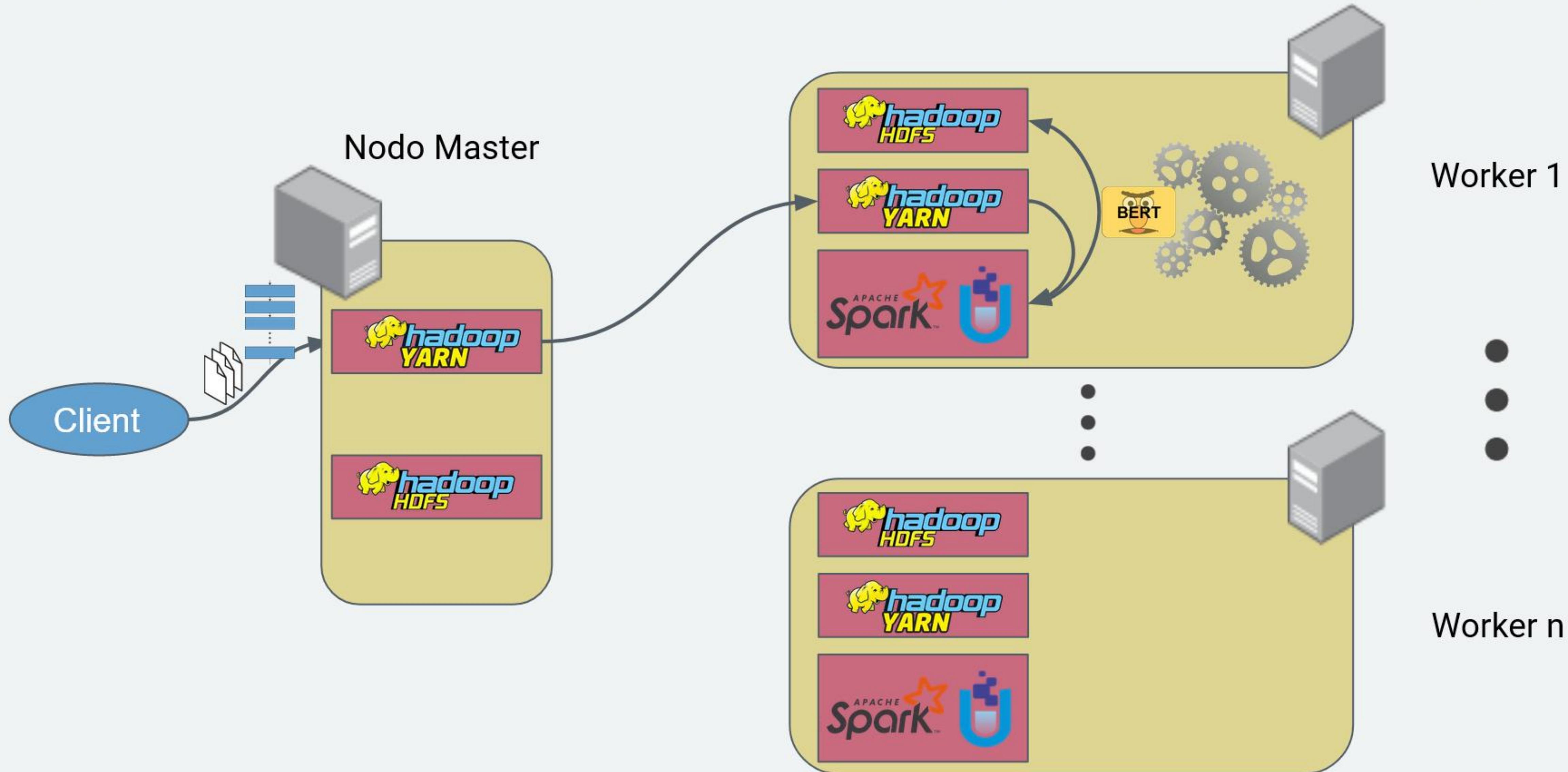
Calcolo distribuito con Spark e Hadoop



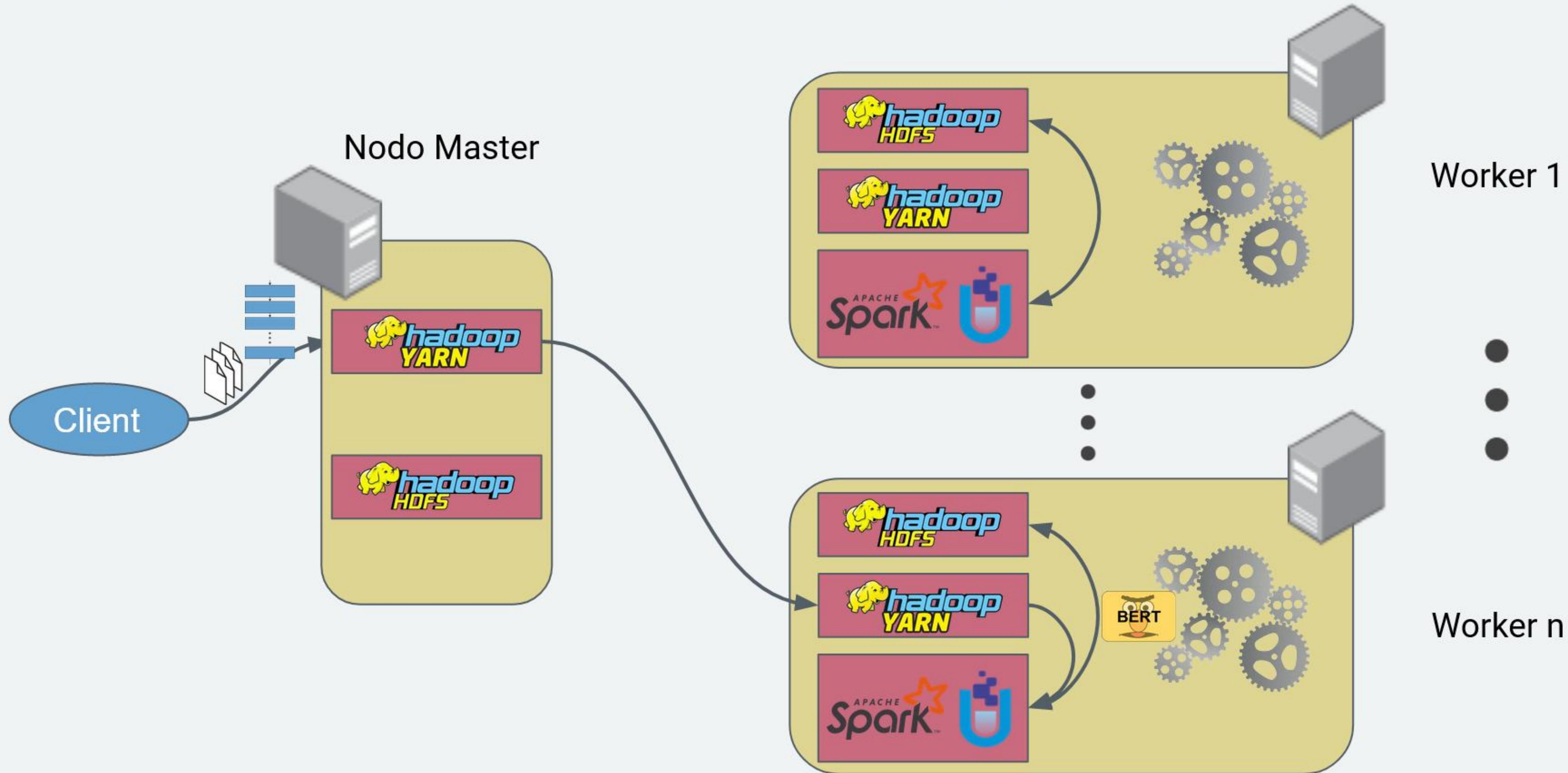
Calcolo distribuito con Spark e Hadoop



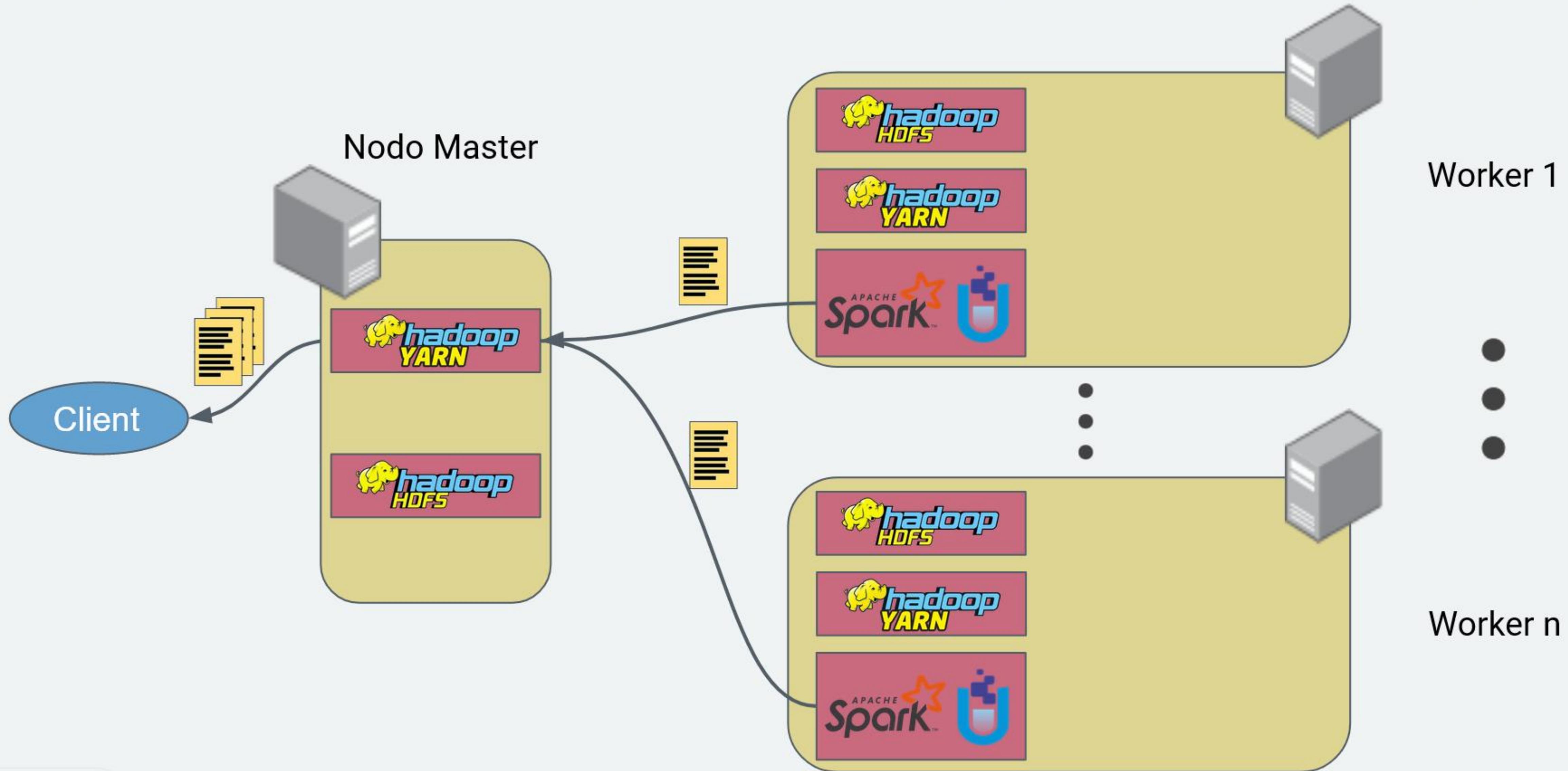
Calcolo distribuito con Spark e Hadoop



Calcolo distribuito con Spark e Hadoop



Calcolo distribuito con Spark e Hadoop



Valutazione sperimentale

Obiettivi della fase sperimentale:

- Misurare l'accuratezza ottenuta utilizzando modelli allo stato dell'arte

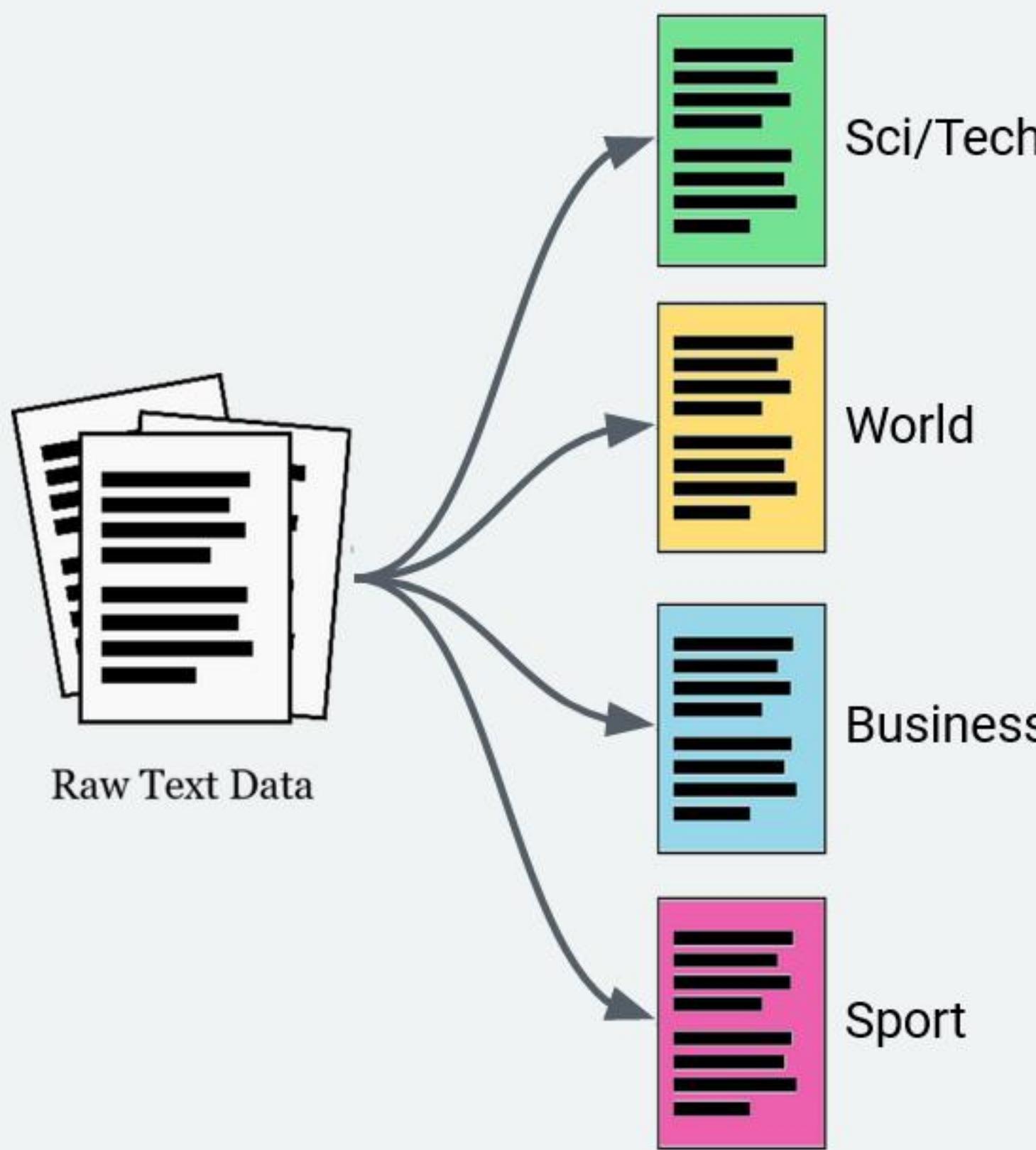


- Misurare la scalabilità offerta da Spark NLP eseguito su di un sistema distribuito.



I Task affrontati

Text Classification (Topic Detection)



Entity Extraction (Named Entity Recognition)

In fact, the Chinese NORP market has the three CARDINAL most influential names of the retail and tech space – Alibaba GPE, Baidu ORG, and Tencent PERSON (collectively touted as BAT ORG), and is betting big in the global AI GPE in retail industry space. The three CARDINAL giants which are claimed to have a cut-throat competition with the U.S. GPE (in terms of resources and capital) are positioning themselves to become the 'future AI PERSON platforms'. The trio is also expanding in other Asian NORP countries and investing heavily in the U.S. GPE based AI GPE startups to leverage the power of AI GPE. Backed by such powerful initiatives and presence of these conglomerates, the market in APAC AI is forecast to be the fastest-growing one CARDINAL, with an anticipated CAGR PERSON of 45% PERCENT over 2018 - 2024 DATE.

To further elaborate on the geographical trends, North America LOC has procured more than 50% PERCENT of the global share in 2017 DATE and has been leading the regional landscape of AI GPE in the retail market. The U.S. GPE has a significant credit in the regional trends with over 65% PERCENT of investments (including M&As, private equity, and venture capital) in artificial intelligence technology. Additionally, the region is a huge hub for startups in tandem with the presence of tech titans, such as Google ORG, IBM ORG, and Microsoft ORG.

Spark NLP per Text Classification

Lettura dataset

Esempi di addestramento:

120.000

Esempi per la valutazione:

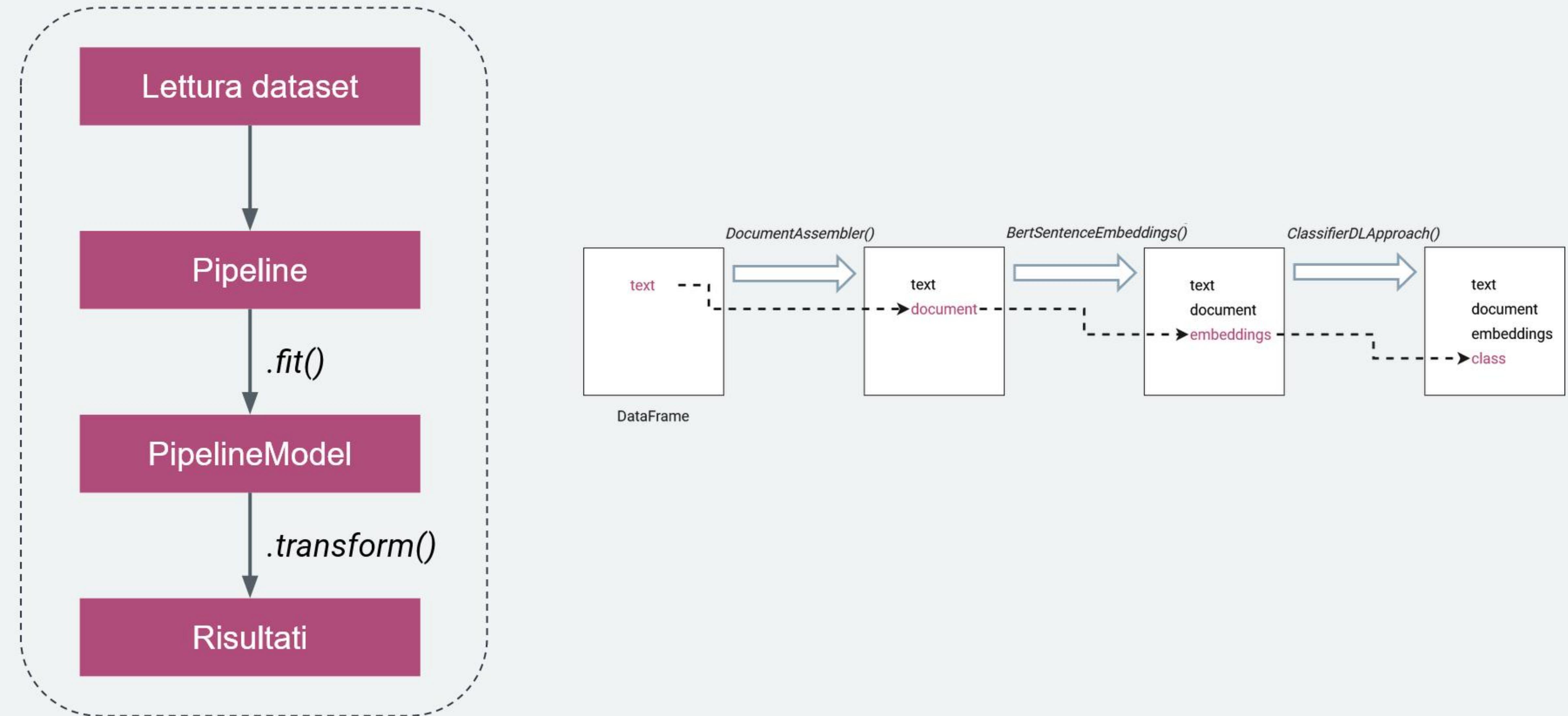
7.600

Classi:

- Business
- Sci/Tech
- World
- Sports

DataFrame

Spark NLP per Text Classification



Risultati ottenuti

- Supponiamo di avere come riferimento un modello che, per ogni esempio presente nel dataset di valutazione, restituisce sempre la stessa classe.
- Nel nostro dataset il numero di esempi per ogni classe è lo stesso, pertanto, otterrebbe una accuracy del 25%
- Il modello utilizzato in questa tesi ottiene i seguenti risultati:

	precision	recall	f1-score	support
Business	0.85	0.86	0.85	7544
Sci/Tech	0.88	0.85	0.87	7868
Sports	0.97	0.94	0.96	7844
World	0.87	0.92	0.90	7144
accuracy			0.89	30400
macro avg	0.89	0.89	0.89	30400
weighted avg	0.90	0.89	0.89	30400

Esempio di errore:

"Supporters and rivals warn of possible fraud; government says Chavez's defeat could produce turmoil in world oil market."

Classe attesa:
'World'

Classe predetta:
'Business'

Risultati ottenuti

- Supponiamo di avere come riferimento un modello che, per ogni esempio presente nel dataset di valutazione, restituisce sempre la stessa classe.
- Nel nostro dataset il numero di esempi per ogni classe è lo stesso, pertanto, otterrebbe una accuracy del 25%
- Il modello utilizzato in questa tesi ottiene i seguenti risultati:

	precision	recall	f1-score	support
Business	0.85	0.86	0.85	7544
Sci/Tech	0.88	0.85	0.87	7868
Sports	0.97	0.94	0.96	7844
World	0.87	0.92	0.90	7144
accuracy			0.89	30400
macro avg	0.89	0.89	0.89	30400
weighted avg	0.90	0.89	0.89	30400

Esempio di errore:

“Richard Faulds and Stephen Parry are going for gold for Great Britain on day four in Athens.”

Classe attesa:
“World”

Classe predetta:
“Sports”

Tempi di esecuzione

Esempi:
760.000

1
esecutore

BERT

8h 6min 03s

Tempi di esecuzione

Esempi:
760.000

BERT

**1
esecutore**

8h 6min 03s

**2
esecutori**

4h 6min 05s



Aggiungendo un esecutore ci aspettiamo che il tempo dimezzi

Le aspettative sono state rispettate!

Conclusioni

Scalabilità

Ottimi risultati:
I tempi di esecuzione
diminuiscono
linearmemente rispetto al
numero di esecutori

Accuratezza

Notevoli risultati
Strumenti e modelli
allo stato dell'arte
facilmente
implementabili

Prospettive

Approfondire le
opzioni fornite dal
framework
Sperimentazione su
sistemi composti da
un numero maggiore
di worker
Utilizzo di modelli
diversi
Sperimentazione su
dati eterogenei

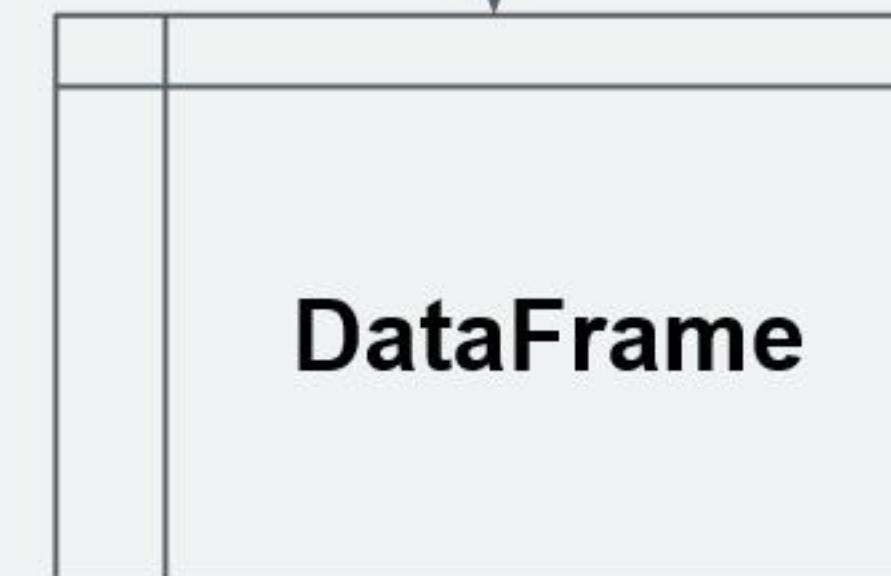
Grazie!

Ci sono domande?

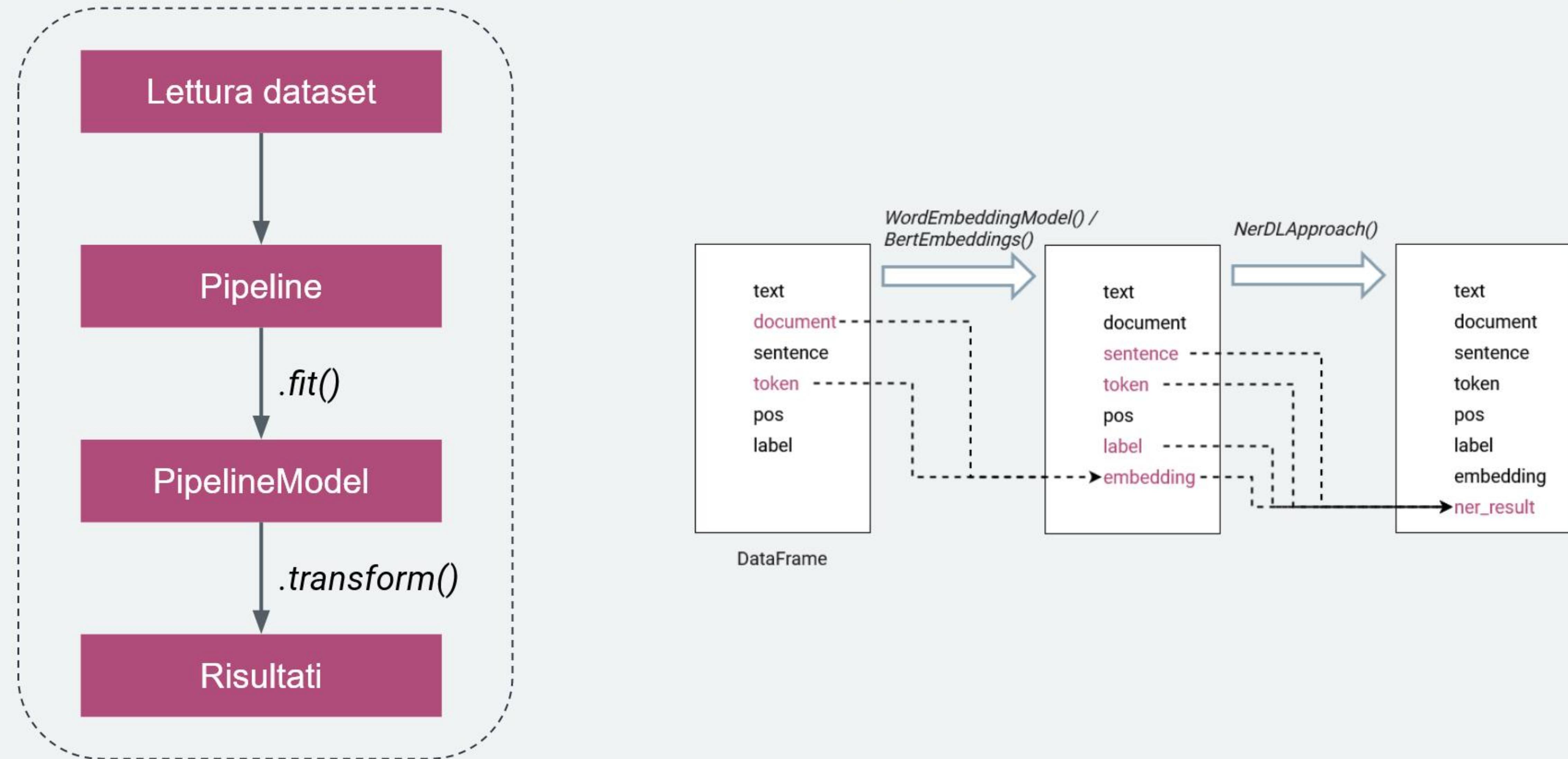
Spark NLP per NER

Lettura dataset

Lingua Italiana	Lingua Inglese
Esempi per l'addestramento: 11.227 Esempi per la valutazione: 4.136 Formato: CoNLL 2003 Entità: <ul style="list-style-type: none">• Persone• Luoghi• Organizzazioni• Entità geopolitiche	Esempi per l'addestramento: 14.041 Esempi per la valutazione: 6.603 Formato: CoNLL 2003 Entità: <ul style="list-style-type: none">• Person• Location• Organization• Miscellaneous



Spark NLP per NER



Risultati ottenuti

GloVe Embeddings

	precision	recall	f1-score	support
LOC	0.89	0.90	0.90	14020
MISC	0.82	0.71	0.76	6496
ORG	0.80	0.79	0.80	12008
PER	0.92	0.95	0.93	13836
micro avg	0.87	0.86	0.86	46360
macro avg	0.86	0.84	0.85	46360
weighted avg	0.87	0.86	0.86	46360

BERT Embeddings

	precision	recall	f1-score	support
LOC	0.81	0.93	0.87	14020
MISC	0.81	0.76	0.79	6496
ORG	0.87	0.75	0.80	12008
PER	0.94	0.95	0.95	13836
micro avg	0.86	0.87	0.87	46360
macro avg	0.86	0.85	0.85	46360
weighted avg	0.87	0.87	0.86	46360

NER Inglese

Risorse

- Template **Slidesgo**
- Icone: **Flaticon**
- Immagini: **Freepik, Pixabay**
- Illustrazioni: **Stories by Freepik**