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

Correlatore:
Prof. Danilo Croce

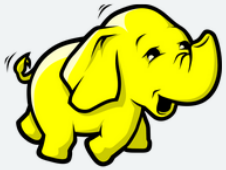
Candidato:
Manuel Di Lullo

Macroarea di Scienze Matematiche, Fisiche e Naturali - Corso di Laurea in Informatica a.a. 2020/2021

Obiettivi della tesi

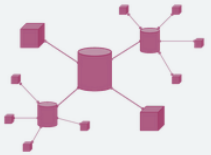
1

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

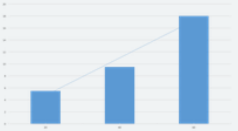


3

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

4

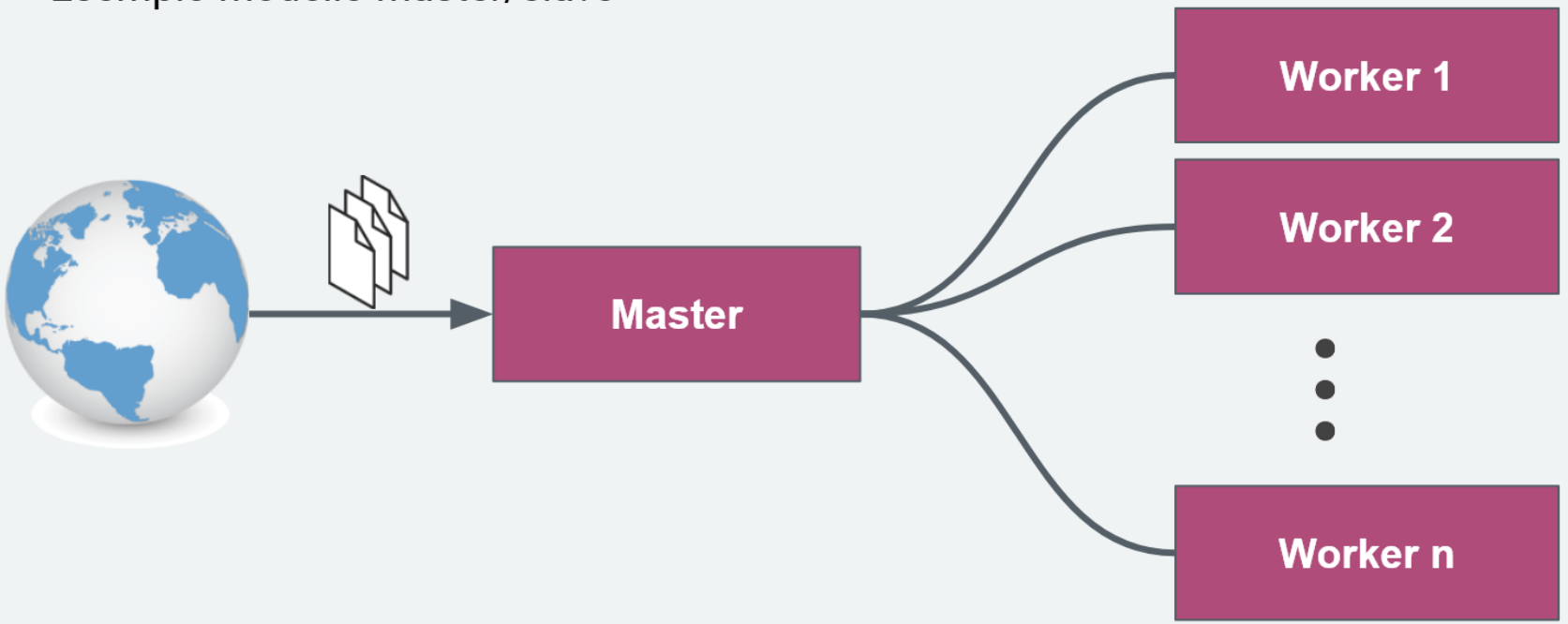
Valutazione sperimentale dei modelli. Misura dell'accuratezza e della scalabilità della soluzione.



Un architettura per 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.

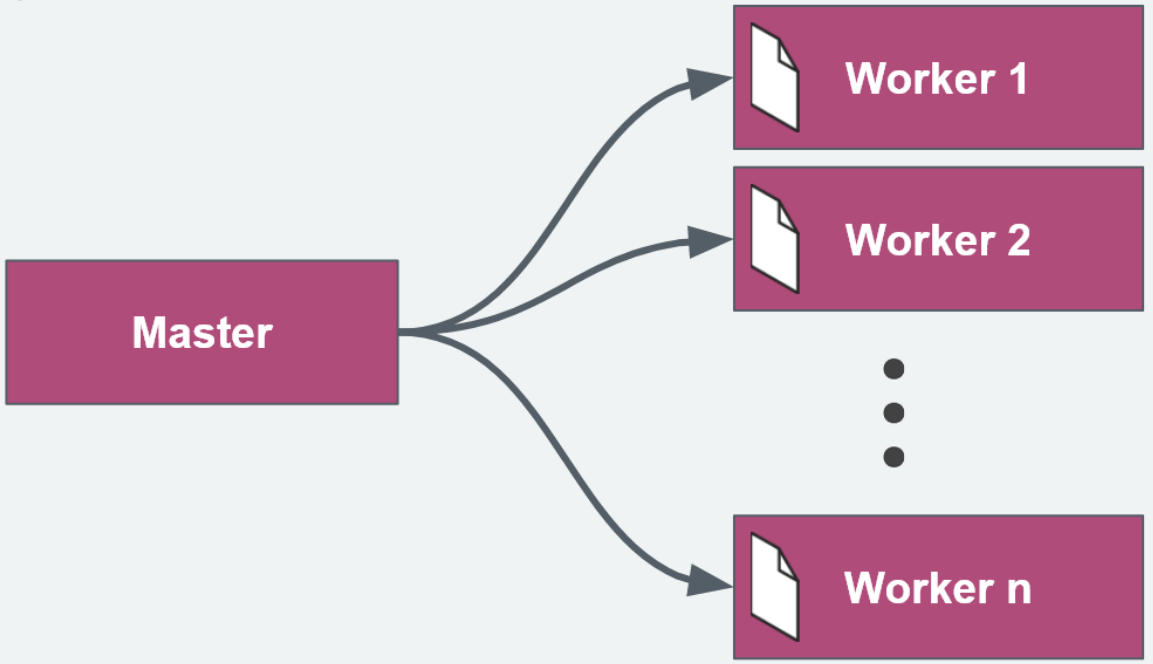
Esempio modello master/slave



Un architettura per 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.

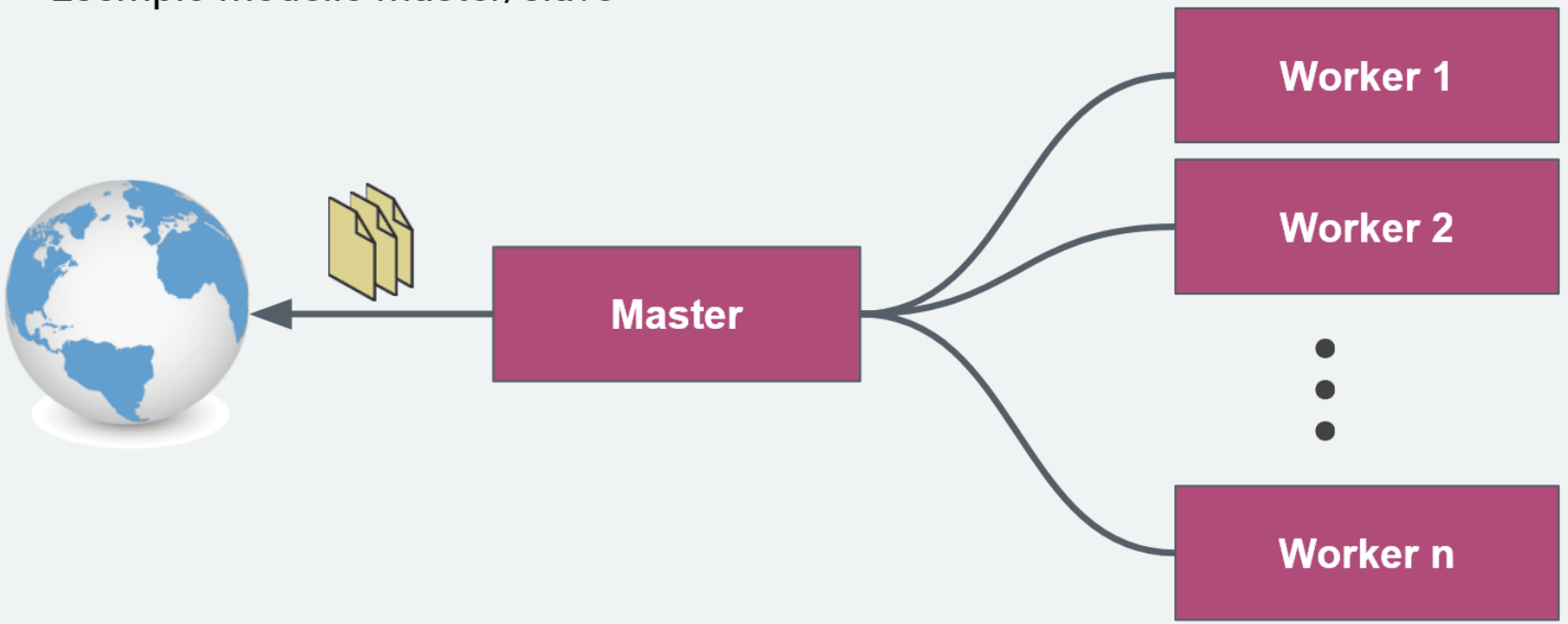
Esempio modello master/slave



Un architettura per 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



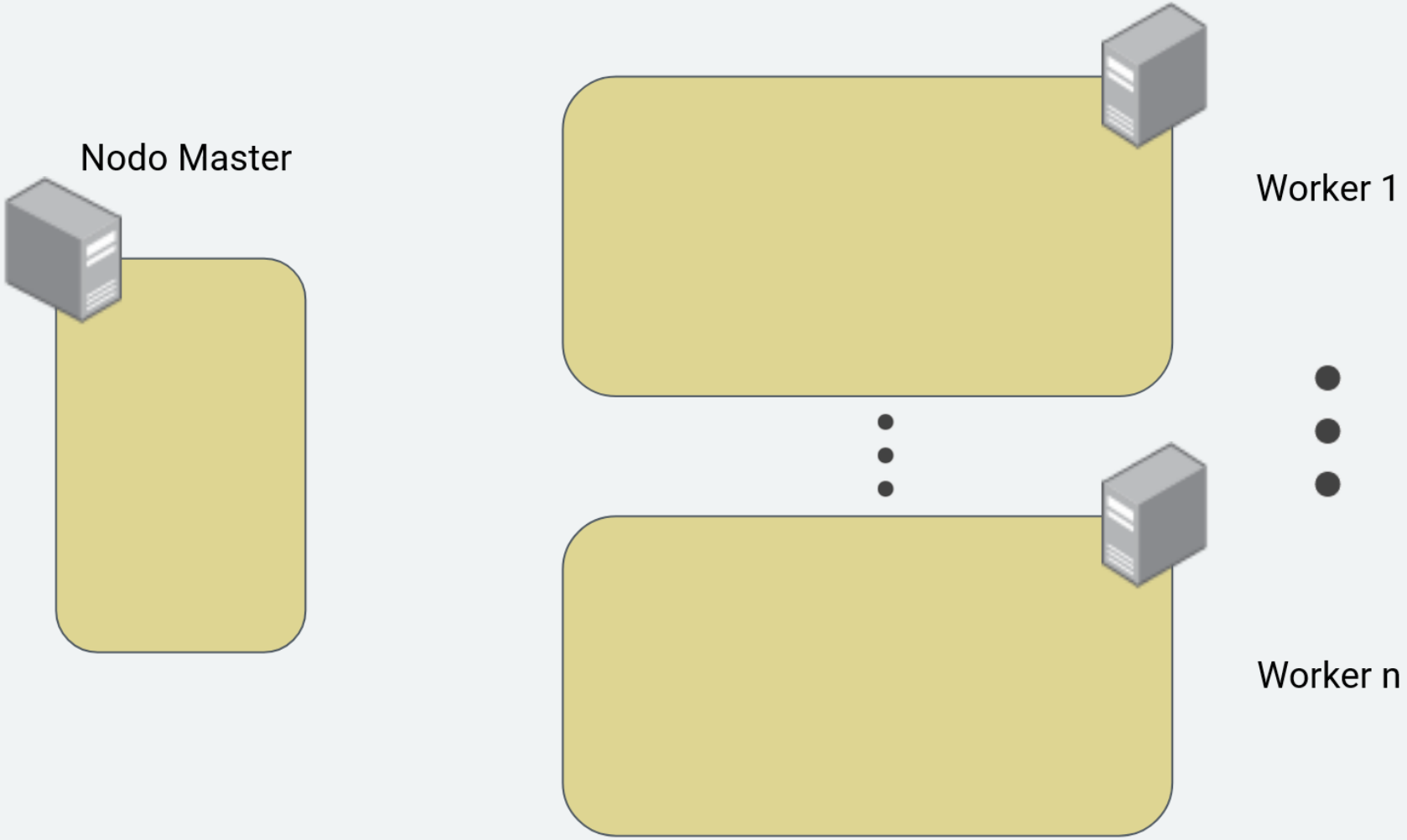
Soluzione proposta

Elaborazione dei dati in ambiente distribuito

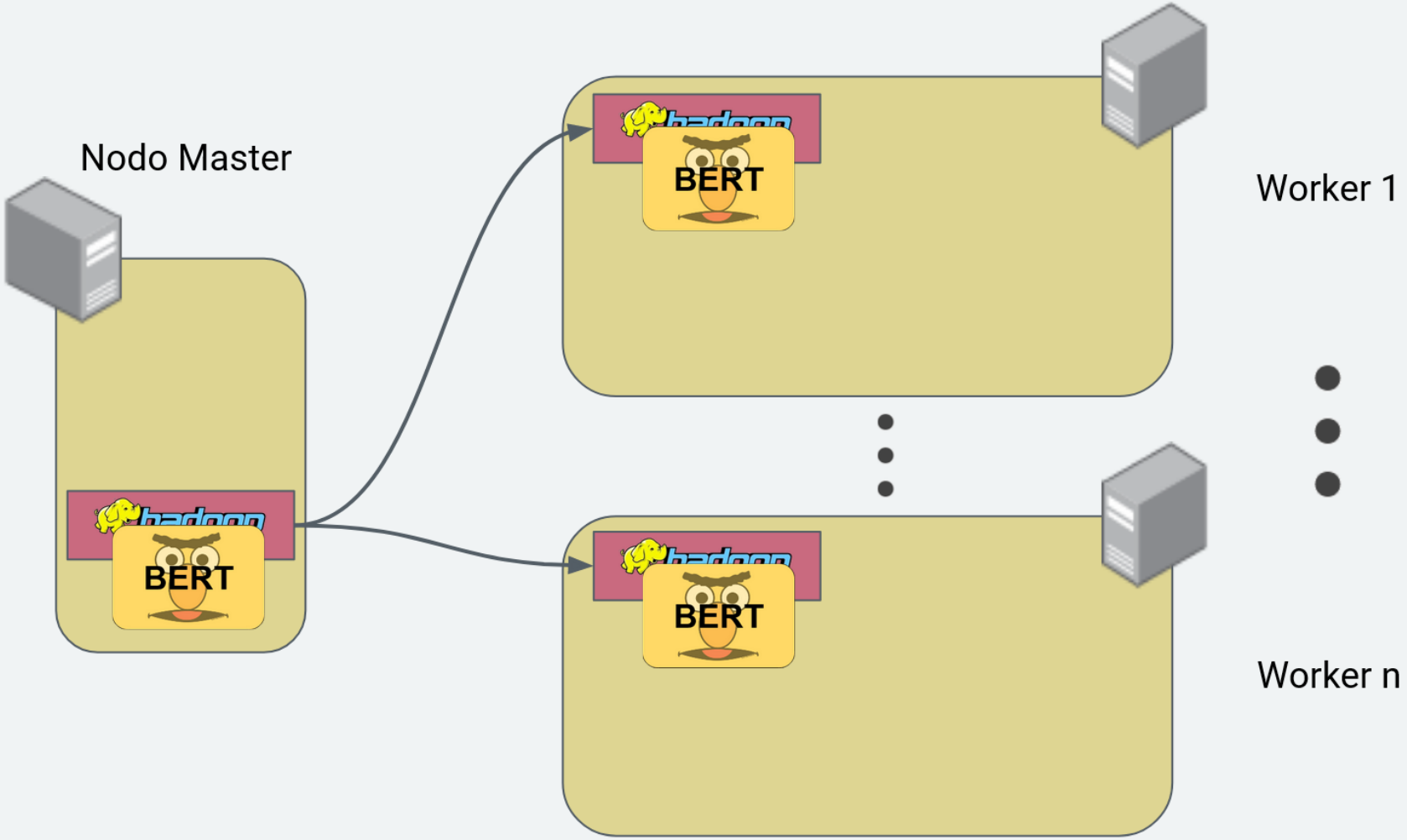
Scheduling e gestione delle risorse

Gestione dei file in ambiente distribuito

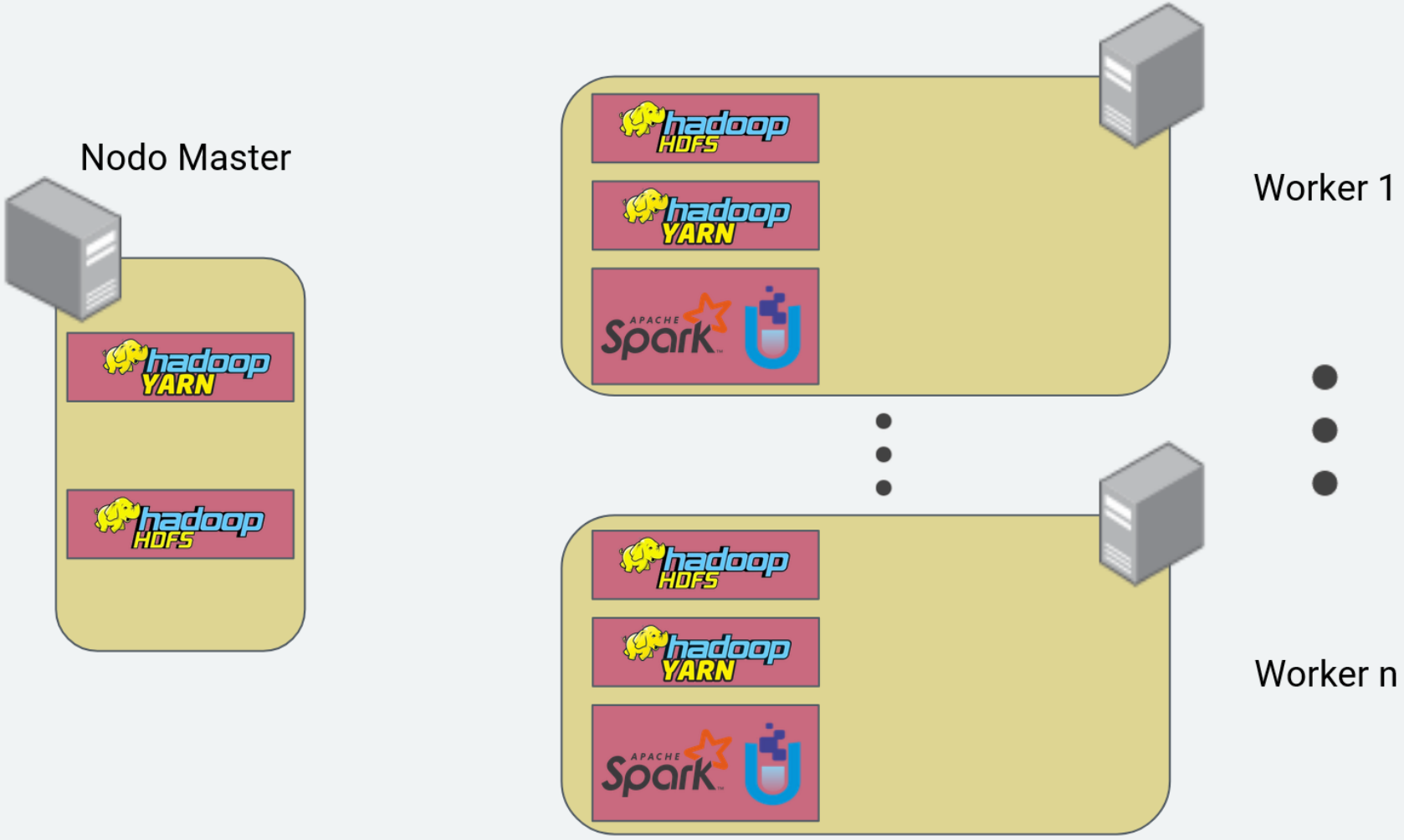
Calcolo distribuito con Spark e Hadoop



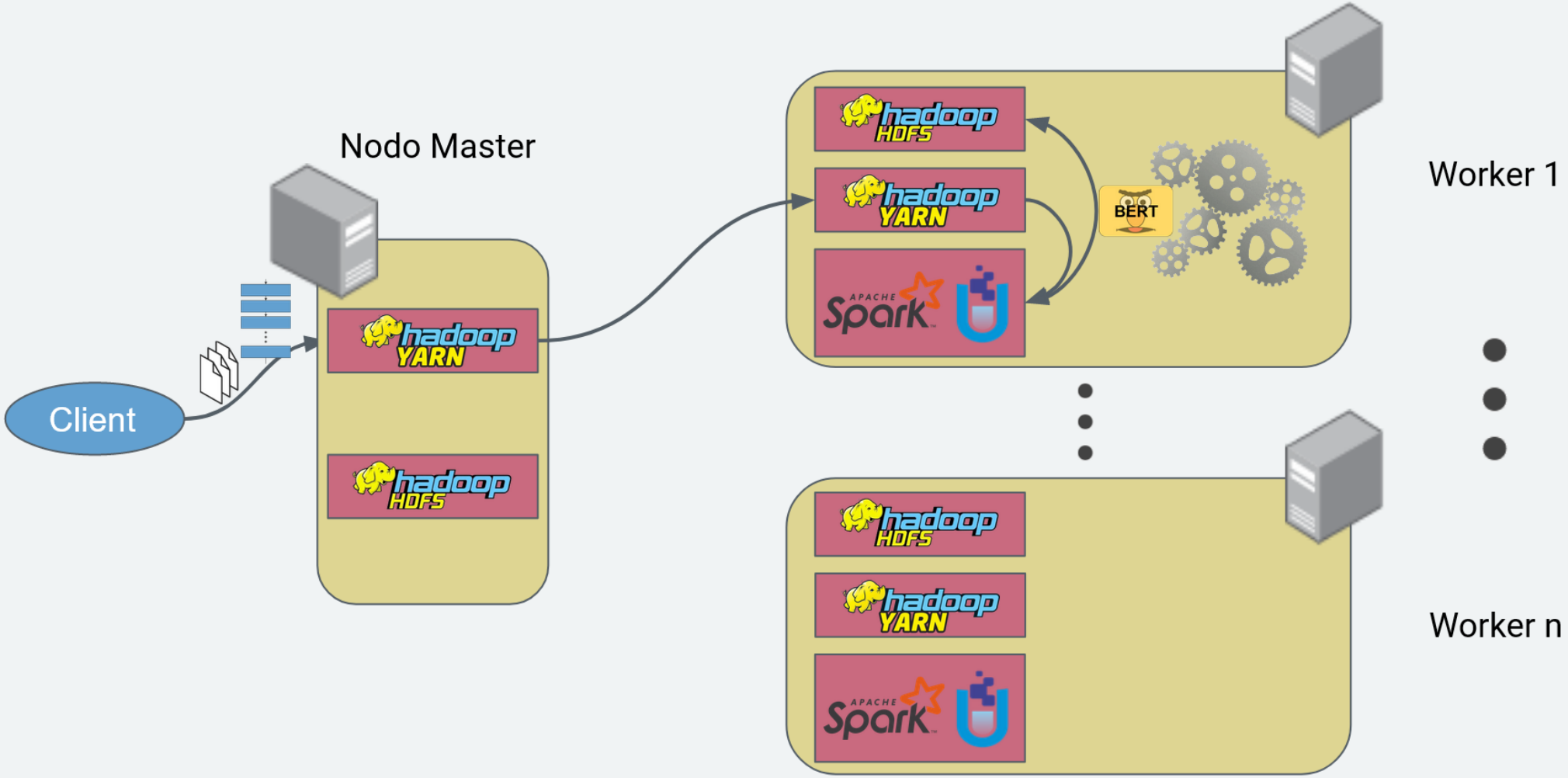
Calcolo distribuito con Spark e Hadoop



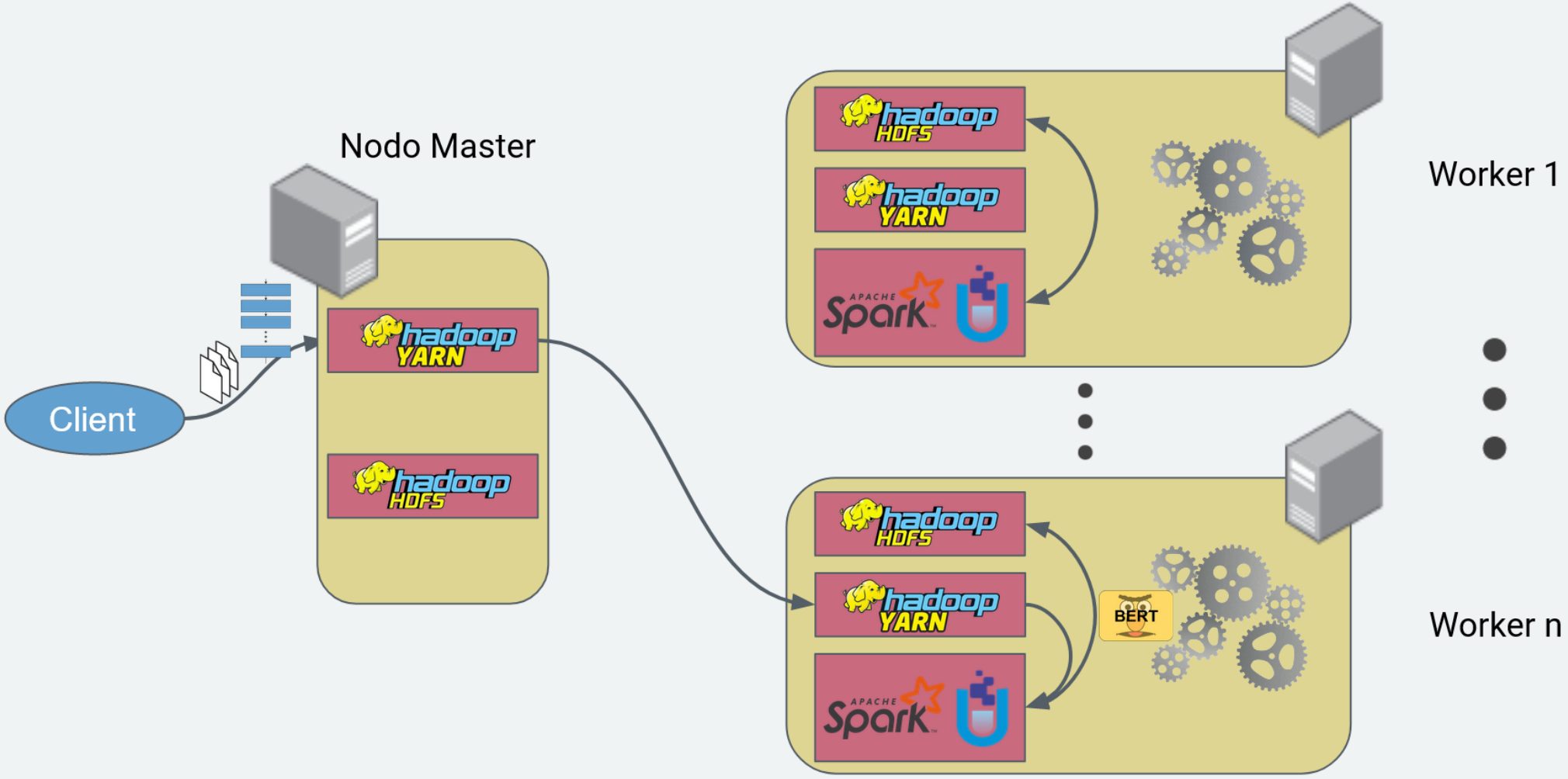
Calcolo distribuito con Spark e Hadoop



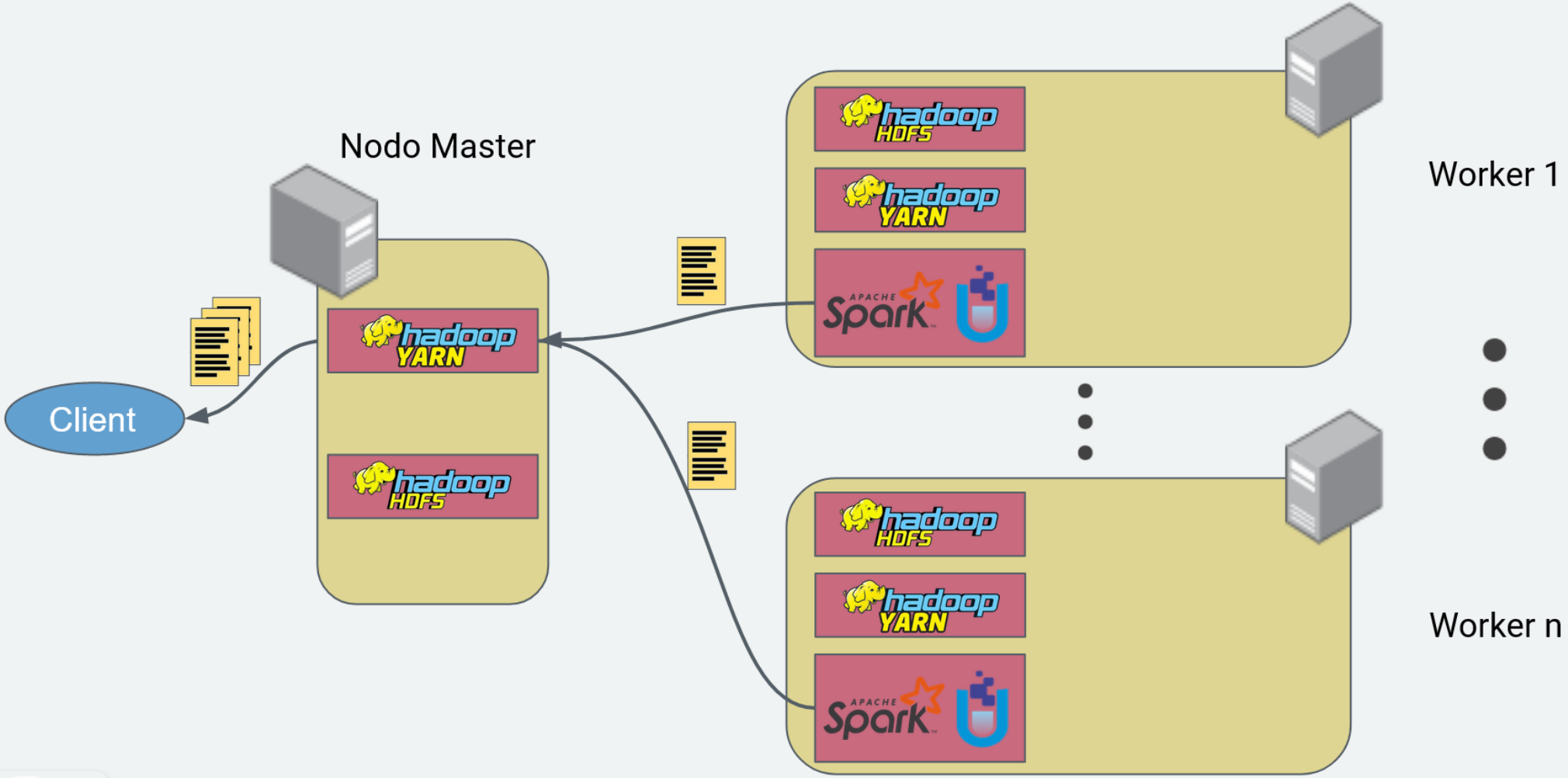
Calcolo distribuito con Spark e Hadoop



Calcolo distribuito con Spark e Hadoop



Calcolo distribuito con Spark e Hadoop



Risultati ottenuti

Per la risoluzione dei task sono stati utilizzati diversi modelli, tra cui: BERT, Universal Sentence Encoder e GloVE

Risultati in termini di accuratezza

Task	Risultato
Text Classification	89%
Named Entity Recognition	87%

Risultati in termini di scalabilità*

# Esempi: 750.000	1 esecutore	2 esecutori
BERT	8h 6min 03s	4h 6min 05s

*Valutazione effettuata soltanto sul task di text classification

Conclusioni

Scalabilità

Ottimi risultati:

I tempi di esecuzione diminuiscono linearmente 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?