



<http://ims.dei.unipd.it/>

Esperienze del DEI con l'utilizzo del Cloud

Dr. Ing. Gianmaria Silvello

Gruppo di ricerca di Sistemi di Gestione dell'Informazione (IMS), Dipartimento di Ingegneria dell'Informazione,
Università degli Studi di Padova

silvello@dei.unipd.it

Ing. Paolo E. Mazzon

Dipartimento di Ingegneria dell'Informazione,
Università degli Studi di Padova

mazzon@dei.unipd.it

Seminario di presentazione e utilizzo della piattaforma Cloud per il calcolo scientifico

25 Novembre 2015



Use Case 1: Motore di ricerca (Java)

- Indicizzazione di collezioni di documenti testuali con dimensioni variabili
 - Caso tipico: 1 milione di documenti con dimensione totale 8 GB
 - Indicizzazione in Java con molte operazioni di lettura e scrittura

- Recupero di documenti interrogando gli indici
 - Per ogni collezione si eseguono (in genere) 50 interrogazioni con diverse configurazioni di sistema
 - Per ogni configurazione si produce un file di risultati che occupa dai 5 ai 10 MB
 - Abbiamo testato 640 combinazioni per 39 collezioni = ~25K file
 - Dimensione totale dell'output prodotto per ogni collezione ~3GB
- Necessità:
 - **Dati condivisi** tra diverse istanze del motore di ricerca
 - Operazioni di recupero in parallelo utilizzando diverse istanze del motore di ricerca
 - **Spazio su disco**

- Per ogni collezione di test :
 - Si importano in Matlab i dati ottenuti dal motore di ricerca producendo una tabella (tipicamente) di dimensioni 50x640, ogni cella contiene un'altra tabella 1000x3 con i dettagli sui documenti recuperati
 - Ogni tabella occupa in memoria centrale ~6GB
 - Le misure calcolate sono memorizzate in tabelle di uguale dimensione che occupano ~2GB ciascuna
- Necessità:
 - Frequenti operazioni di lettura e scrittura su disco
 - **Molto spazio in RAM** (4 o 5 tabelle di dati in memoria più lo spazio necessario per l'elaborazione dei dati e a contenere le tabelle con le misure)

- I test condotti con un motore di ricerca Java hanno dato buoni frutti
 - Esecuzioni parallele di indicizzazioni e recupero documenti

- I test condotti con Matlab hanno rivelato la necessità di disporre di un maggior quantitativo di memoria centrale
 - Esecuzioni parallele limitate dalla dimensione della memoria



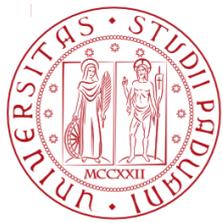
Un primo test

Cloud, istanza “large” con 4 VCPU e 8GB RAM

VS

Hardware fisico, 1 blade Dell PowerEdge M610,
Intel Xeon six-core X5680 (24 core totali), 48 GB RAM

Nota: la macchina fisica ha 6x core e 6x RAM



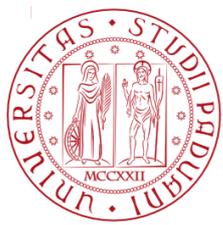
Un primo test: prestazioni Java

- Indicizzazione documenti
 - **Cloud: 282.9 seconds**
 - HW: 469 seconds



Un primo test: prestazioni Matlab

- Importazione e creazione tabelle (50x116) di dati in memoria:
 - Cloud: 9 minutes, 28.1016 seconds
 - **HW: 5 minutes, 15.6712 seconds**



Un primo test: prestazioni Matlab

- Calcolo di una tabella di misure (50x116) in memoria:
 - Cloud: 12 minutes, 51.0441 seconds
 - **HW: 5 minutes, 43.5721 seconds**

- Le prestazioni Java sono confrontabili se non superiori
- *Per quanto riguarda Matlab l'istanza in cloud non ha 1/6 delle prestazioni.*
 - Matlab e' un software che viene molto ottimizzato sull'hardware sottostante.
 - Nella cloud il processore viene «virtualizzato» per motivi tecnici
- Il test si potrebbe rifare creando una istanza di pari dimensioni (*o facendo il resize di quella esistente*)

- Le nostre necessità principali sono:
 - Disporre di grandi quantità di dati in memoria di massa condivisibili tra più istanze del cloud.
 - Scrivere e leggere frequentemente file da memoria di massa.
 - Disporre di grandi quantitativi di memoria centrale per l'elaborazione dei dati.

- Fare un upgrade della piattaforma HW significherebbe
 - Acquistare una nuova blade
 - Metterla in rete con quelle già esistenti
 - Reinstallare da zero tutto il software
 - Utilizzare un server per condividere i dati che non è a nostro uso esclusivo

COSTI

- Piu' di 1K euro
- svariate ore uomo

- Fare un upgrade nella piattaforma **Cloud** significherebbe
 - Fare il resize di una istanza già esistente, oppure
 - Clonare una istanza già esistente
 - Nessuna configurazione di rete
 - Tutte le istanze utilizzano una loro rete esclusiva
 - Nessuna reinstallazione del software
 - Poter utilizzare un server per condividere i dati ad uso esclusivo del progetto

COSTI:

- qualche decina di click del mouse
- riutilizzo delle ore uomo impiegate finora

- La cloud si presenta come una tecnologia «abilitante», permette di sperimentare configurazioni diverse in maniera flessibile
- Le possibilità' di modificare a piacimento le risorse di calcolo non hanno paragoni con l'hw fisico
- Molte possibilità' sono ancora da esplorare...

Grazie per l'attenzione