P.O.N. RICERCA E COMPETITIVITA' 2007-2013 - Azione I "Interventi di rafforzamento strutturale"
PONA3_00052, Avv. 254/Ric. - ReCaS (Rete di Calcolo per SuperB e altre applicazioni)


ABSTRACT sottoposto in risposta alla Call for Papers, sul tema "operations di grandi infrastrutture di calcolo distribuito"

# Titolo: ***Sviluppo, implementazione e testing di un'infrastruttura di storage distribuito basata su Hadoop-FS per la gestione dei dati per calcolo scientifico ad alta affidabilità in centri di calcolo distribuiti***

*Autori*: ***D. Diacono, G. Donvito, on behalf of the ReCaS collaboration, e G. Marzulli, Consortium GARR e INFN Bari***

## Abstract:

Nel presente lavoro verrà mostrata l'attività svolta per implementare e testare una installazione distribuita di Hadoop-FS (HDFS) fra diversi siti INFN che partecipano al PON potenziamento Infrastrutturale ReCaS.
Il file-system HDFS è generalmente usato come software di gestione dati all'interno di una singola farm. Questo lavoro ha avuto come obbiettivo quello di cambiare alcune componenti in modo da poterlo usare con successo in un ambiente geograficamente distribuito.
In particolare verranno mostrate le modifiche effettuate alle regole di "data-placement" e di "data-replication" che sono disponibili in HDFS.
HDFS, infatti, è nativamente capace di decidere la localizzazione e la replica dei files in base alla disposizione di data-nodes in racks, per garantire che anche la "failure" di un intero rack non comporti la perdita di dati nel file-system.
HDFS ha nativamente anche la possibilità di esprimere la composizione del data center con racks a diversi livelli di aggregazione; tali livelli però normalmente non vengono sfruttati ed ogni rack viene considerato singolarmente indipendentemente dal livello a cui si trova.

Con le modifiche applicate al codice di HDFS, durante questo lavoro, è ora possibile scegliere la localizzazione dei dati in fase di scrittura considerando la gerarchia e la topologia dei "racks".
Tali funzionalità si applicano in due contesti diversi:

- All'interno di una farm locale possono essere utilizzate per comporre diverse "gerarchie di fallimento". Questa granularità è di fondamentale aiuto quando si ha l'esigenza di fornire una elevata tolleranza ai guasti dei vari componenti dell'infrastruttura di base (in rack, su presiere di alimentazione, su switch, etc).
- Nella distribuzione dei dati su più farm geograficamente distribuite, in modo da conservare la disponibilità del dato anche in caso di perdita di un intero centro di calcolo. Questa funzionalità consente di distribuire, in fase di scrittura, i dati in modo da avere due copie nella farm locale e una copia in una farm remota. Questo meccanismo garantisce che i dati siano comunque disponibili, all'interno della farm sorgente, anche in caso di perdita di un intero rack. In fase di lettura ovviamente viene preferito l'accesso alla repliche locali e solo in caso queste non siano disponibili il dato viene letto dalla farm remota.

In questo lavoro si mostreranno i risultati del test effettuati in un "testbed" costituito dalle farm di INFN-Bari e INFN-Napoli per verificare il corretto funzionamento delle policy di scrittura e lettura.
Verrà mostrata, inoltre, l'attività di verifica delle performance di scrittura e lettura in una farm di calcolo configurata con 210 nodi di calcolo e circa 130TB di storage, sia con standard benchmark che con job reali di analisi da parte di utenti di fisica e di bioinformatica.
Oltre ai test delle nuove funzionalità e delle relative performance, saranno descritte le attività portate avanti allo scopo di verificare l'integrazione del file-system HDFS in una farm grid/cloud in produzione (INFN-Bari).

Le principali attività sono di seguito elencate:

- Configurazione del sistema di monitoring attualmente usato nella farm, basato su Ganglia, per controllare lo stato dei processi di HDFS.
- Sviluppo di un sistema di installazione e configurazione automatica di HDFS. Questo consente la messa in produzione di centinaia di nodi senza lunghi e noiosi interventi umani.
- Sviluppo e messa in produzione di un sistema aggiuntivo di monitoring della localizzazione dei blocchi. Tale sistema, implementato con un RDBMS, permette di sapere su quali dischi ognuno dei blocchi è stato registrato. Sono anche conservate le informazioni storiche che possono essere utili in fase di risoluzione di eventuali problemi.
- Verifica della possibilità di usare HDFS con autenticazione basata su Kerberos.
- Messa a punto di una procedura per la compilazione dei moduli di FUSE che risultano necessari al fine di utilizzare il file-system con una interfaccia posix-like.

Referenze:


1 HADOOP: http://hadoop.apache.org
2 FUSE: http://fuse.sourceforge.net/
3 Kerberos: http://web.mit.edu/kerberos/