Introduzione

Questo libro vi guiderà per mano nell'esplorazione del fantastico mondo di Hadoop 2 e del suo ecosistema in continua crescita. Basato sulle solide fondamenta delle versioni precedenti della piattaforma, Hadoop 2 consente l'esecuzione di più framework di elaborazione dei dati su un unico cluster Hadoop.
Per darvi un'idea della portata dell'evoluzione, studieremo sia il funzionamento dei nuovi modelli sia come si applicano nell'elaborazione di grandi volumi di dati con algoritmi batch, iterativi e quasi in tempo reale.

Struttura del libro

Il Capitolo 1, "Per iniziare", fornisce le basi di Hadoop e per affrontare i problemi dei big data che intende risolvere. Vedremo anche dove Hadoop 1 ha spazio di miglioramento.
Il Capitolo 2, "Storage", entra nel dettaglio dell'Hadoop Distributed File System, dove vengono memorizzati i dati elaborati da Hadoop. Esamineremo le caratteristiche specifiche di HDFS, spiegheremo come utilizzarle e vedremo come è migliorato in Hadoop 2. Presenteremo anche ZooKeeper, un altro sistema di storage all'interno di Hadoop, su cui si basano molte funzionalità cruciali.
Il Capitolo 3, "Elaborazione: MapReduce e oltre", affronta innanzitutto il tradizionale modello di elaborazione di Hadoop e come viene utilizzato. Vedremo poi come Hadoop 2 ha generalizzato la piattaforma per utilizzare più modelli computazionali, tra i quali MapReduce è solo uno dei tanti.
Il Capitolo 4, "Computazione in tempo reale con Samza", approfondisce uno di questi modelli di elaborazione alternativi abilitati da Hadoop 2. In particolare, vedremo come elaborare dati in streaming in tempo reale con Apache Samza.
Il Capitolo 5, "Computazione iterativa con Spark", entra nel merito di un modello di elaborazione molto diverso. In questo capitolo parleremo dei mezzi forniti da Apache Spark per effettuare l'elaborazione iterativa.

Il Capitolo 6, "Analisi dei dati con Apache Pig", mostra come Apache Pig semplifichi l'uso del modello computazionale tradizionale di MapReduce attraverso un linguaggio che descrive i flussi di dati.
Il Capitolo 7, "Hadoop e SQL", analizza come il familiare linguaggio SQL è stato implementato sui dati salvati in Hadoop. Attraverso l'uso di Apache Hive e la descrizione di alternative come Cloudera Impala, vedremo come rendere possibile l'elaborazione dei big data usando le competenze e gli strumenti esistenti.
Il Capitolo 8, "Gestione del ciclo di vita dei dati", dà un'occhiata generale a come gestire tutti i dati che devono essere elaborati in Hadoop. Attraverso Apache Oozie, illustreremo come costruire dei workflow per ottenere, elaborare e gestire i dati.
Il Capitolo 9, "Facilitare il lavoro di sviluppo", si concentra sulla scelta degli strumenti che devono aiutare lo sviluppatore a raggiungere rapidamente dei risultati. Attraverso Hadoop Streaming, Apache Crunch e Kite, vedremo come l'uso dello strumento giusto può velocizzare il ciclo di sviluppo o fornire nuove API con una semantica più ricca e meno ridondanze.
Il Capitolo 10, "Eseguire un cluster Hadoop", considera il lato operativo di Hadoop. Concentrandosi sugli ambiti di primo interesse degli sviluppatori, come la gestione dei cluster, il monitoraggio e la sicurezza, questo capitolo vi aiuterà a lavorare meglio con il vostro staff *operations*.
Il Capitolo 11, "Come proseguire", vi guida in un breve tour tra alcuni progetti e strumenti che riteniamo utili ma che non possiamo trattare nel dettaglio per ragioni di spazio. Vi forniremo anche alcune indicazioni su dove trovare altre informazioni e come unirvi alle varie community open source.

Cosa serve per questo libro

Considerato che poche persone dispongono di una serie di computer di scorta, useremo la macchina virtuale Cloudera QuickStart per la maggior parte degli esempi del libro. È un'immagine di una macchina su cui è preinstallato un cluster Hadoop completo. Può essere eseguita su qualsiasi macchina host che supporta VMware o la tecnologia di virtualizzazione VirtualBox.
Esploreremo anche la piattaforma *Amazon Web Services* (AWS) e vedremo come eseguire alcune delle tecnologie Hadoop sul servizio AWS Elastic MapReduce. I servizi AWS sono gestibili attraverso un browser web o un'interfaccia Linux a riga di comando.

Lo scopo del libro

Questo libro è rivolto perlopiù a sviluppatori di sistemi e applicativi che vogliono imparare a risolvere problemi pratici usando il framework Hadoop e i relativi componenti. Nonostante mostreremo gli esempi in linguaggi di programmazione diversi, il requisito fondamentale è una conoscenza solida di Java. Gli ingegneri e

gli architetti dei dati troveranno utile il materiale riguardante il ciclo di vita dei dati, i formati dei file e i modelli computazionali.

Convenzioni

In questo libro abbiamo applicato stili di testo diversi per distinguere i vari tipi di informazioni. Ecco alcuni esempi con una spiegazione del loro significato.
Le parti del codice e i nomi dei file sono resi con un carattere `monospaziato`. Gli elementi dell'interfaccia sono invece in *corsivo*, così come i nomi delle directory e le parole nuove o importanti.
Un blocco di codice appare così:

```
topic_edges_grouped = FOREACH topic_edges_grouped {
  GENERATE
    group.topic_id as topic,
    group.source_id as source,
    topic_edges.(destination_id,w) as edges;
}
```

Gli input da riga di comando o l'output appaiono così:

```
$ hdfs dfs -put target/elephant-bird-pig-4.5.jar hdfs:///jar/
$ hdfs dfs –put target/elephant-bird-hadoop-compat-4.5.jar hdfs:///jar/
$ hdfs dfs –put elephant-bird-core-4.5.jar hdfs:///jar/
```

NOTA

Note, suggerimenti e avvertimenti appaiono in questa forma.

Codice degli esempi

Il codice sorgente del libro si trova su GitHub all'indirizzo `https://github.com/learninghadoop2/book-examples`. Gli autori applicheranno le eventuali correzioni al codice e lo manterranno aggiornato di pari passo con l'evoluzione della tecnologia.