

Sommario

- ▶ Gli strumenti Open Source
 - ▶ HDFS
 - ▶ Hadoop
 - ▶ Hive
 - ▶ HBase
 - ▶ Solr (Carrot)
 - ▶ Indexing (Batch ed NRT)
- ▶ Possibili approcci ed architetture
 - ▶ Analisi Batch
 - ▶ Ricerca Full Text
- ▶ Cloudera (con breve DEMO)
- ▶ Conclusioni



HDFS

Introduzione

- ▶ **HDFS (Hadoop Distributed File System)** è un file system distribuito progettato per eseguire su commodity hardware
- ▶ HDFS fornisce alto throughput per le applicazioni che utilizzano enormi dataset



Presupposti e scopo

▶ **Fallimenti Hardware**

- ▶ Il fallimento dei nodi hardware è la norma come l'eccezione

▶ **Enormi Dataset**

▶ **Accesso Streaming ai dati**

- ▶ Alto throughput preferito a bassa latenza

▶ **Modello Simple Coherency**

- ▶ Modello di accesso ai file “write-once-read-many”

▶ **Muovere la computazione è più conveniente che muovere i dati**

▶ **Portabilità fra piattaforme hardware e software eterogenee**

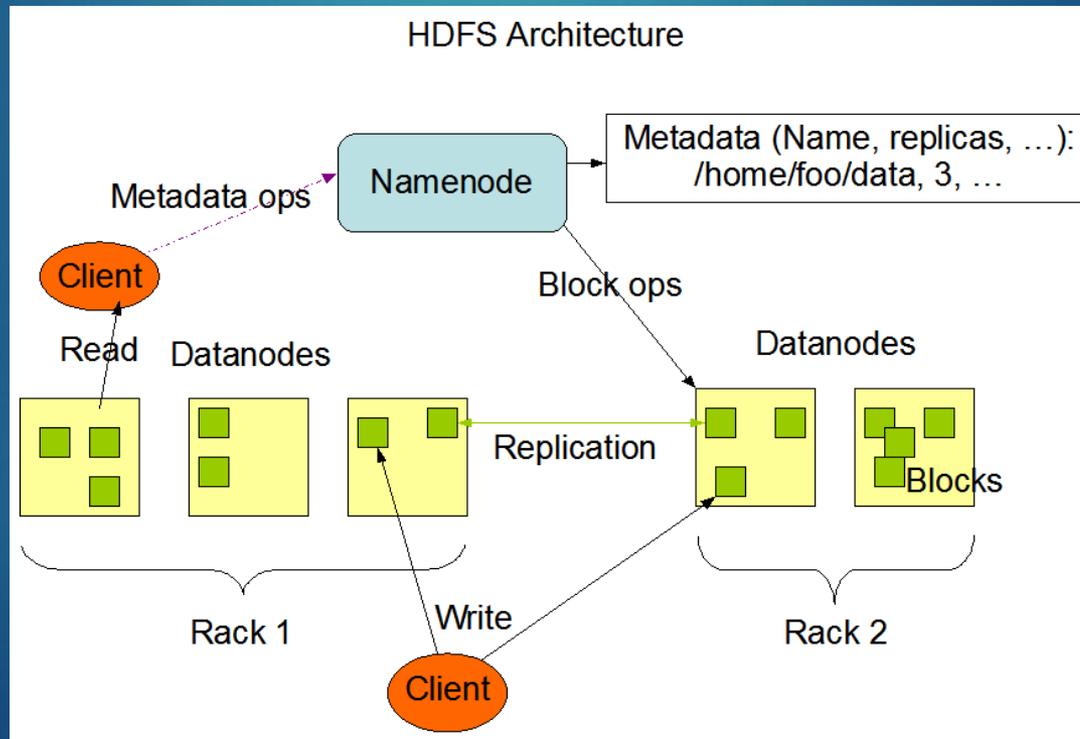
Filesystem Namespace e Architettura

► Namespace

- Organizzazione dei files gerarchica tradizionale

► Architettura Master-Slave

- NameNode & DataNode



Accessibilità

- ▶ **Java API**
- ▶ **C wrapper API**
- ▶ **Fs shell**
 - ▶ HDFS fornisce una interfaccia da linea di comando chiamata FS shell
 - ▶ La sintassi di questi comandi è molto simile a quella dell'altre shell POSIX

Azione	Comando
Creazione di una directory chiamata foodir	> bin/hdfs dfs -mkdir /foodir
Rimozione della directory foodir	> bin/hdfs dfs -rmr /foodir
Visualizzazione del contenuto del file /foodir/Myfile.txt	> bin/hdfs dfs -cat /foodir/myfile.txt

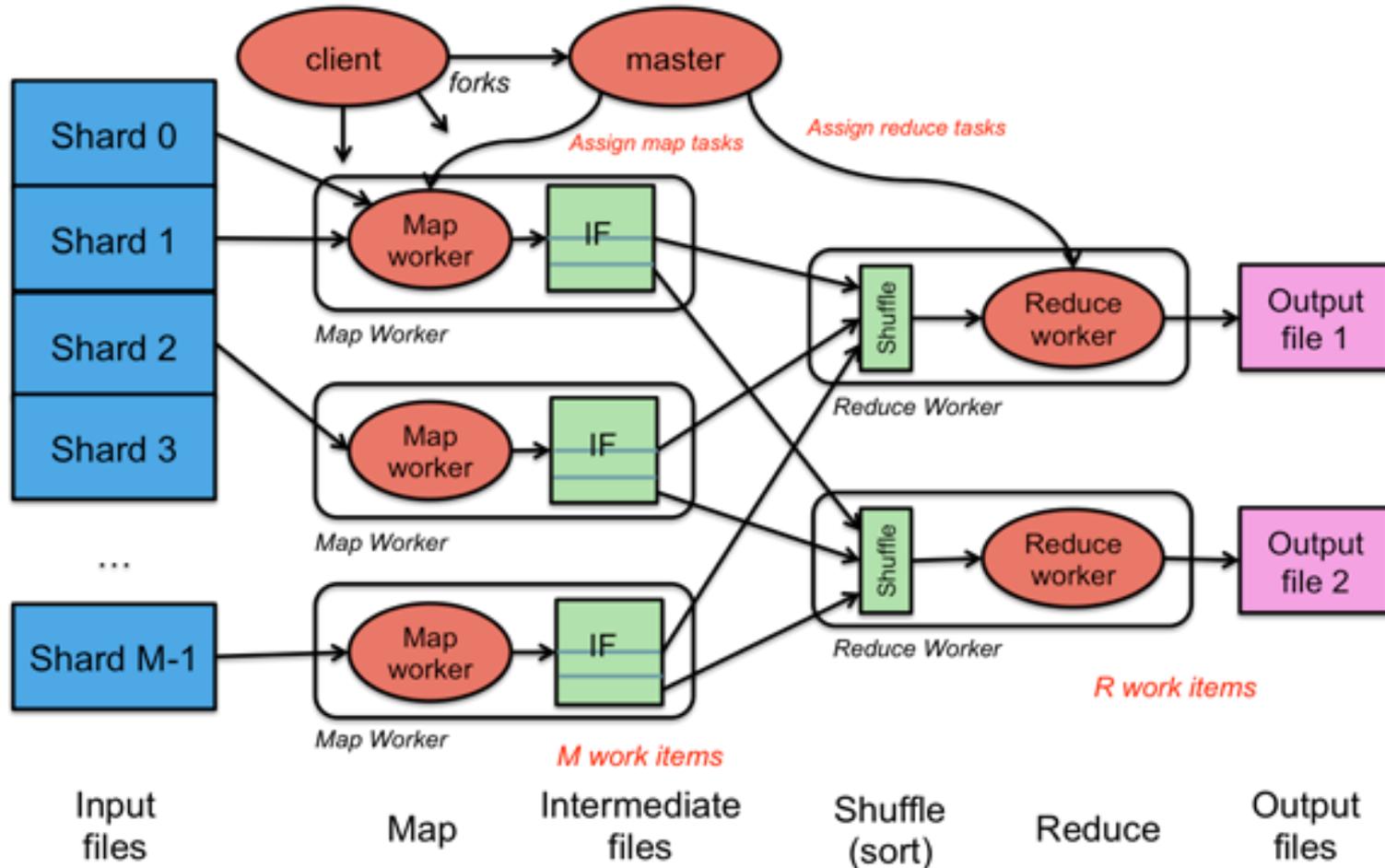


Map-Reduce

Introduzione

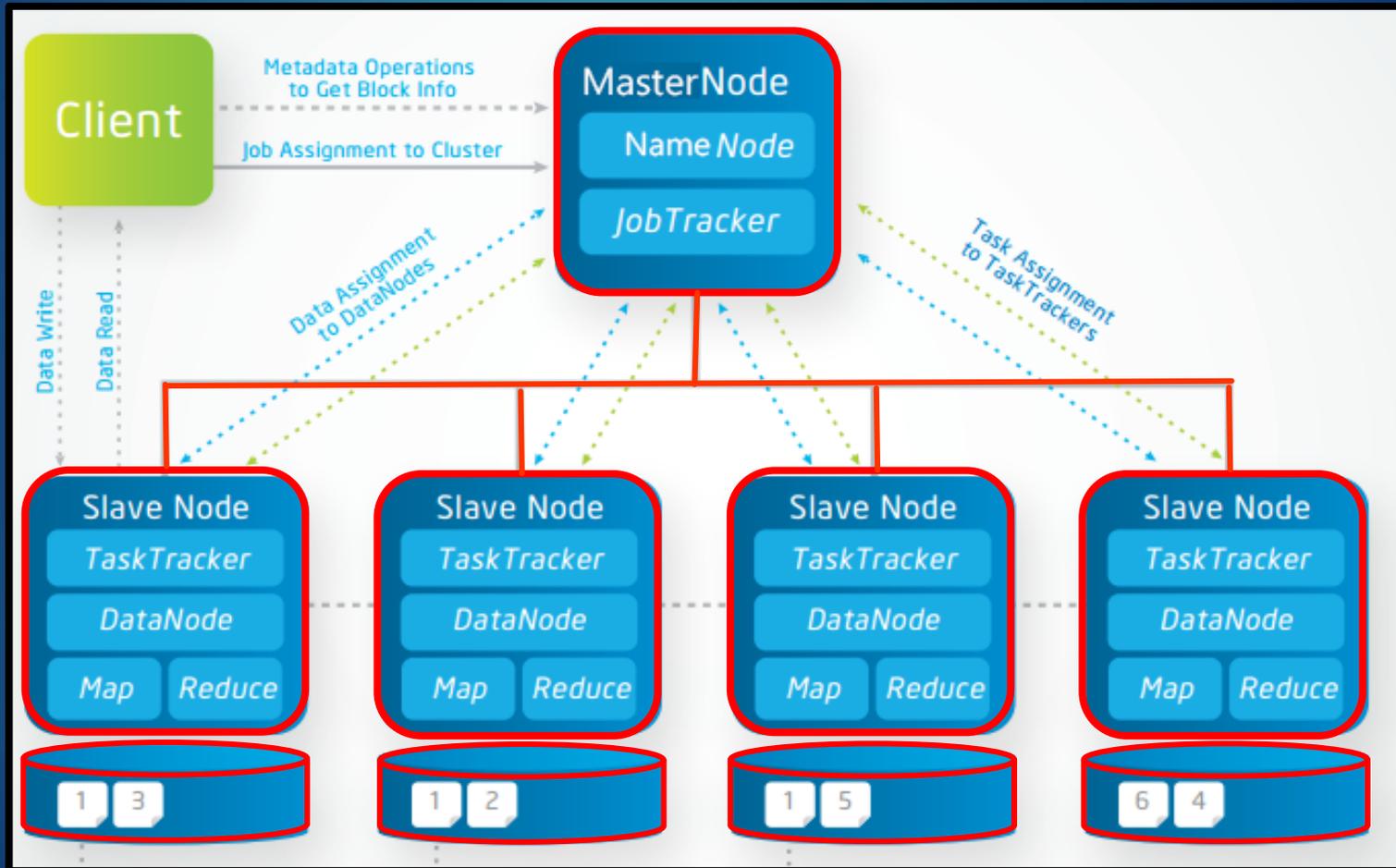
- ▶ **Map-Reduce** è un modello di programmazione, proposto da Google nel 2006, utile a processare e/o generare grandi data sets
- ▶ Gli utenti devono solo specificare una funzione “map” (che processa coppie chiave-valore e ne genera di altre intermedie), ed una funzione “reduce” (che fa il merge di tutti quei valori intermedi associati alla stessa chiave)
- ▶ Rappresenta una soluzione per molti problemi reali e inoltre i programmi scritti secondo questo modello, risultano automaticamente eseguibili in parallelo (anche su clusters di commodity hardware)

MapReduce - Funzionamento

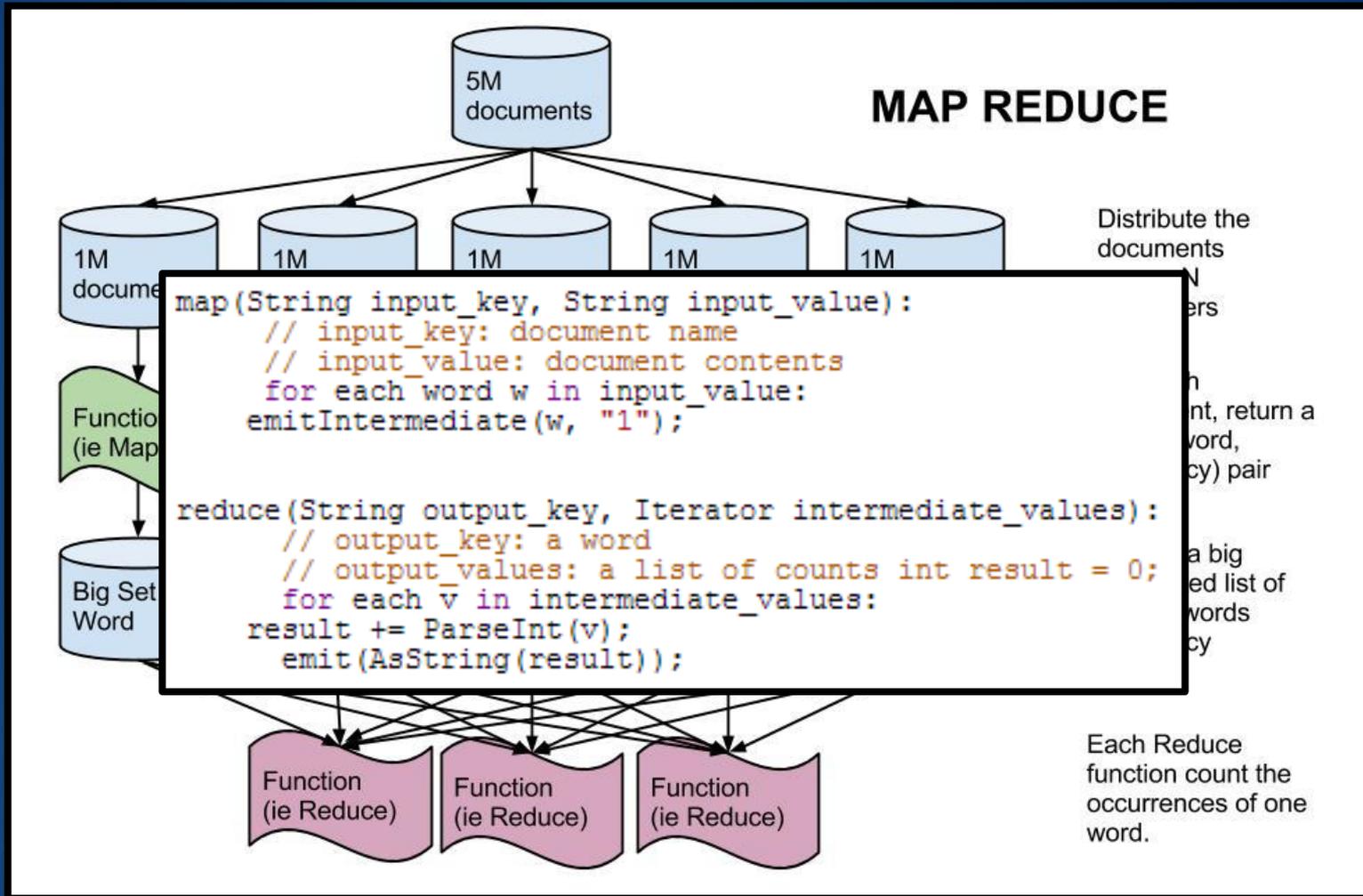


Apache Hadoop

Una implementazione **Open Source** del framework MapReduce



MapReduce – Esempio WordCount

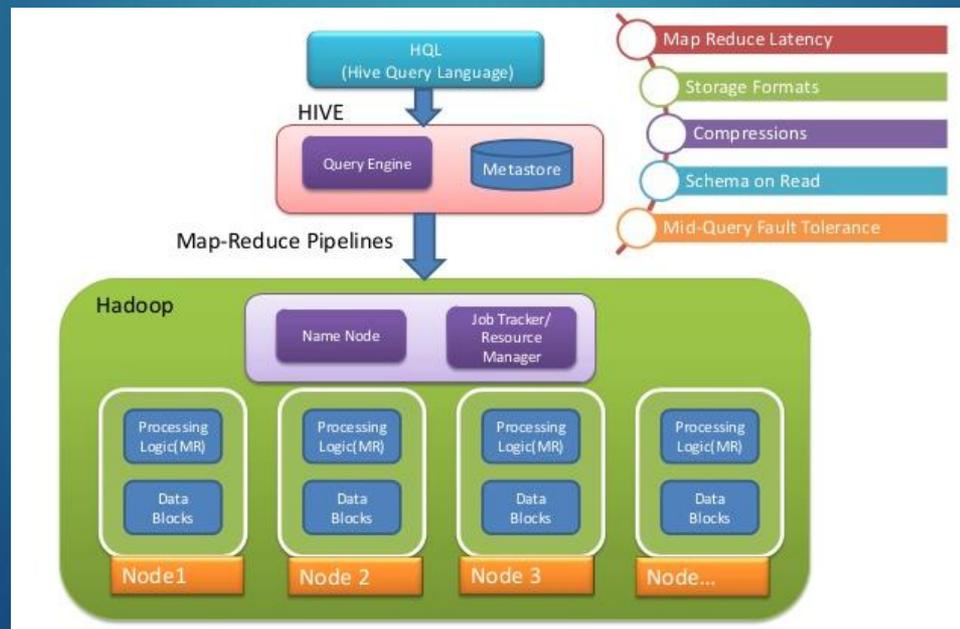




Hive

Introduzione

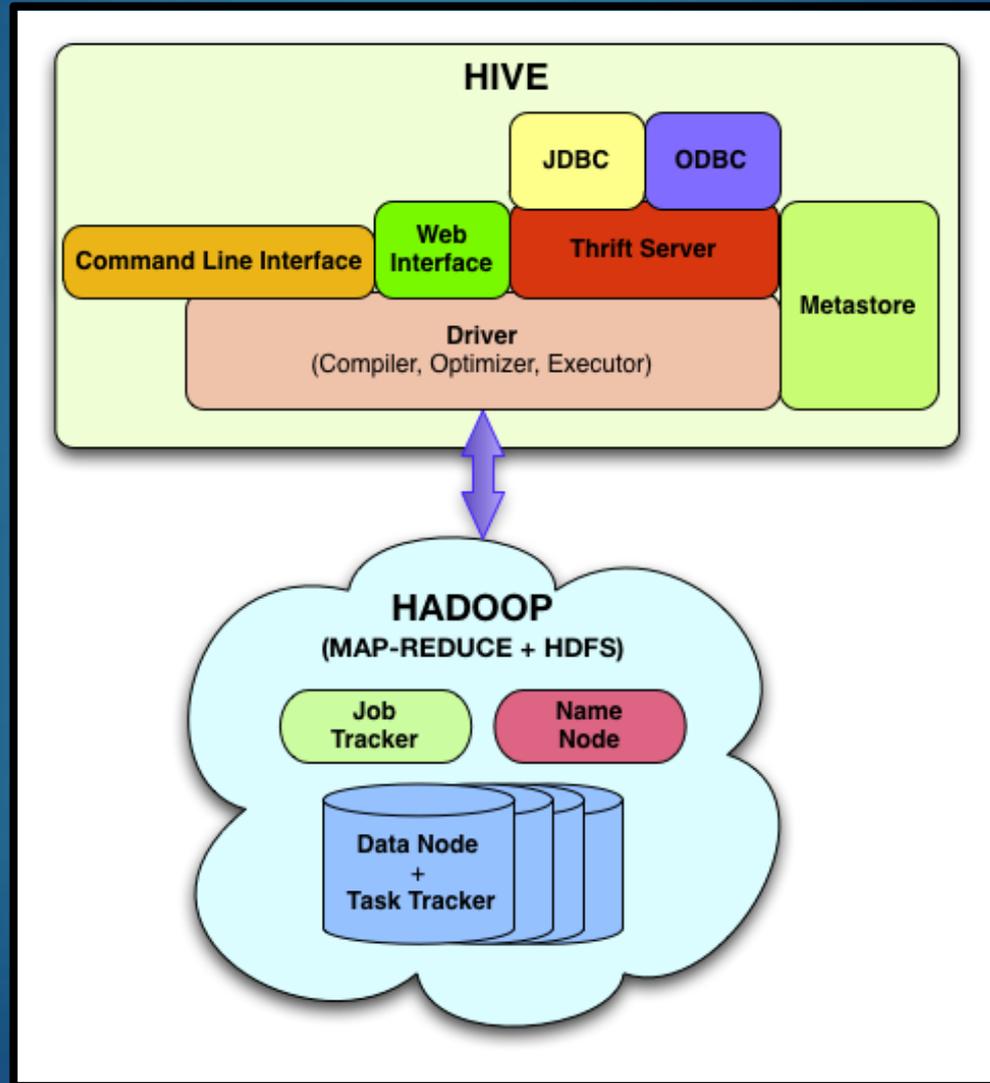
- ▶ **Apache Hive** è un infrastruttura per Data Warehousing creata su Apache Hadoop per fornire dati aggregati, query e analisi
- ▶ Hive fornisce un meccanismo per creare delle strutture sui dati ed effettuare delle query utilizzando un linguaggio SQL-like chiamato HiveQL
- ▶ Trasparentemente all'utente converte le query HiveQL in job map-reduce



Caratteristiche

- ▶ **Tipi di storage**
 - ▶ Plain Text on HDFS
 - ▶ RCFile
 - ▶ HBase
 - ▶ ORC
- ▶ Supporto agli indici
- ▶ Metadati memorizzati su RDMS riducono il tempo di esecuzione di check semantici
- ▶ Operazioni sui dati compressi tramite algoritmi DEFLATE, bwt, snappy

Architettura



HiveQL

▶ SQL

Current SQL Compatibility

Hive SQL Datatypes	Hive SQL Semantics	Color Key
INT	SELECT, LOAD INSERT from query	Hive 0.10
TINYINT/SMALLINT/BIGINT	Expressions in WHERE and HAVING	Hive 0.11
BOOLEAN	GROUP BY, ORDER BY, SORT BY	FUTURE
FLOAT	Sub-queries in FROM clause	
DOUBLE	GROUP BY, ORDER BY	
STRING	CLUSTER BY, DISTRIBUTE BY	
TIMESTAMP	ROLLUP and CUBE	
BINARY	UNION	
ARRAY, MAP, STRUCT, UNION	LEFT, RIGHT and FULL INNER/OUTER JOIN	
DECIMAL	CROSS JOIN, LEFT SEMI JOIN	
CHAR	Windowing functions (OVER, RANK, etc)	
CARCHAR	INTERSECT, EXCEPT, UNION, DISTINCT	
DATE	Sub-queries in WHERE (IN, NOT IN, EXISTS/ NOT EXISTS)	
	Sub-queries in HAVING	

▶ DML

▶ LOAD

▶ INSERT into Hive tables from queries

▶ DDL

▶ CREATE

Hive Caso d'Uso: Facebook

facebook

Hive & Hadoop Usage @ Facebook

- Types of Applications:
 - Reporting
 - Eg: Daily/Weekly aggregations of impression/click counts
 - Measures of user engagement
 - Microstrategy dashboards
 - Ad hoc Analysis
 - Eg: how many group admins broken down by state/country
 - Machine Learning (Assembling training data)
 - Ad Optimization
 - Eg: User Engagement as a function of user attributes
 - Many others

Hive Caso d'Uso: Facebook

facebook

Hadoop & Hive Cluster @ Facebook

- **Production Cluster**
 - 300 nodes/2400 cores
 - 3PB of raw storage
- **Adhoc Cluster**
 - 1200 nodes/9600 cores
 - 12PB of raw storage
- **Node (DataNode + TaskTracker) configuration**
 - 2CPU, 4 core per cpu
 - 12 x 1TB disk (900GB usable per disk)



HBase

Column Family Stores

Permettono di memorizzare i dati su un cluster, partizionandoli sia orizzontalmente che verticalmente. Il partizionamento orizzontale (sharding) delle tabelle, avviene in base alla chiave, mentre le colonne sono distribuite sulla base di un predefinito raggruppamento

- ▶ Essenzialmente sono mappe multi-livello
- ▶ Per accedere ad una cella: `get('row', 'column', 'timestamp')`
- ▶ I valori di una singola colonna sono memorizzati in maniera contigua

Row based

Record 1

Alice	3	25	Bob	4
19	Carol	0	45	

Record 3

Column based

Column A

Alice	Bob	Carol	3		
4	0	25	19	45	

Column C

Column Family based

Column Family A=Column A

Alice	Bob	Carol			
3	25	4	19	0	45

Column Family B = {B,C}

Apache HBase

- ▶ Basato su Big Table (soluzione proposta da Google)
- ▶ Le colonne di una tabella sono raggruppate in Column Families (CFs)
- ▶ Le CFs sono statiche e definite in anticipo
- ▶ Entro la stessa CF, differenti righe possono avere differenti colonne
- ▶ Esiste un solo tipo di dato: il Byte-array

<i>Row id</i>	<i>Column families</i>			
	ColumnFamily1	ColumnFamily2	...	ColumnFamilyN
row_id_1	column1="value1" column2="value2"	column3="value3"
row_id_2				
...				
row_id_l				
row_id_m				

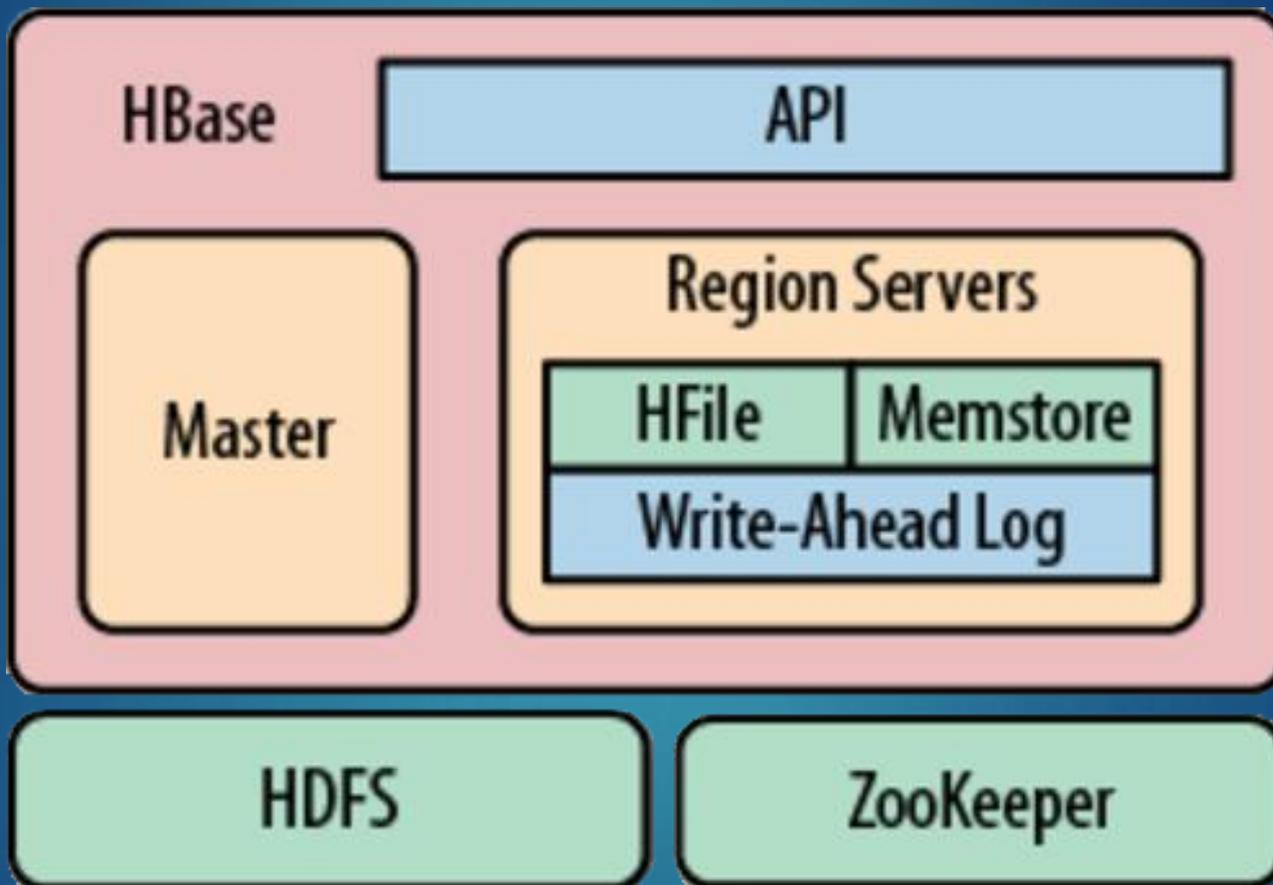
regioni



Operazioni:

- get
- put
- delete
- scan

L'architettura di HBase



Gestione di HBase

- ▶ HBase Shell
- ▶ API Java Native (soluzione più veloce)
- ▶ REST Server (poco efficiente)
- ▶ Thrift Server (compatibile con diversi linguaggi)

Esiste un alto livello di integrazione con il Map Reduce

Permette di memorizzare miliardi di tuple, tuttavia rispetto ai classici RDBMS non è adattabile ad ogni situazione, per via delle API limitate



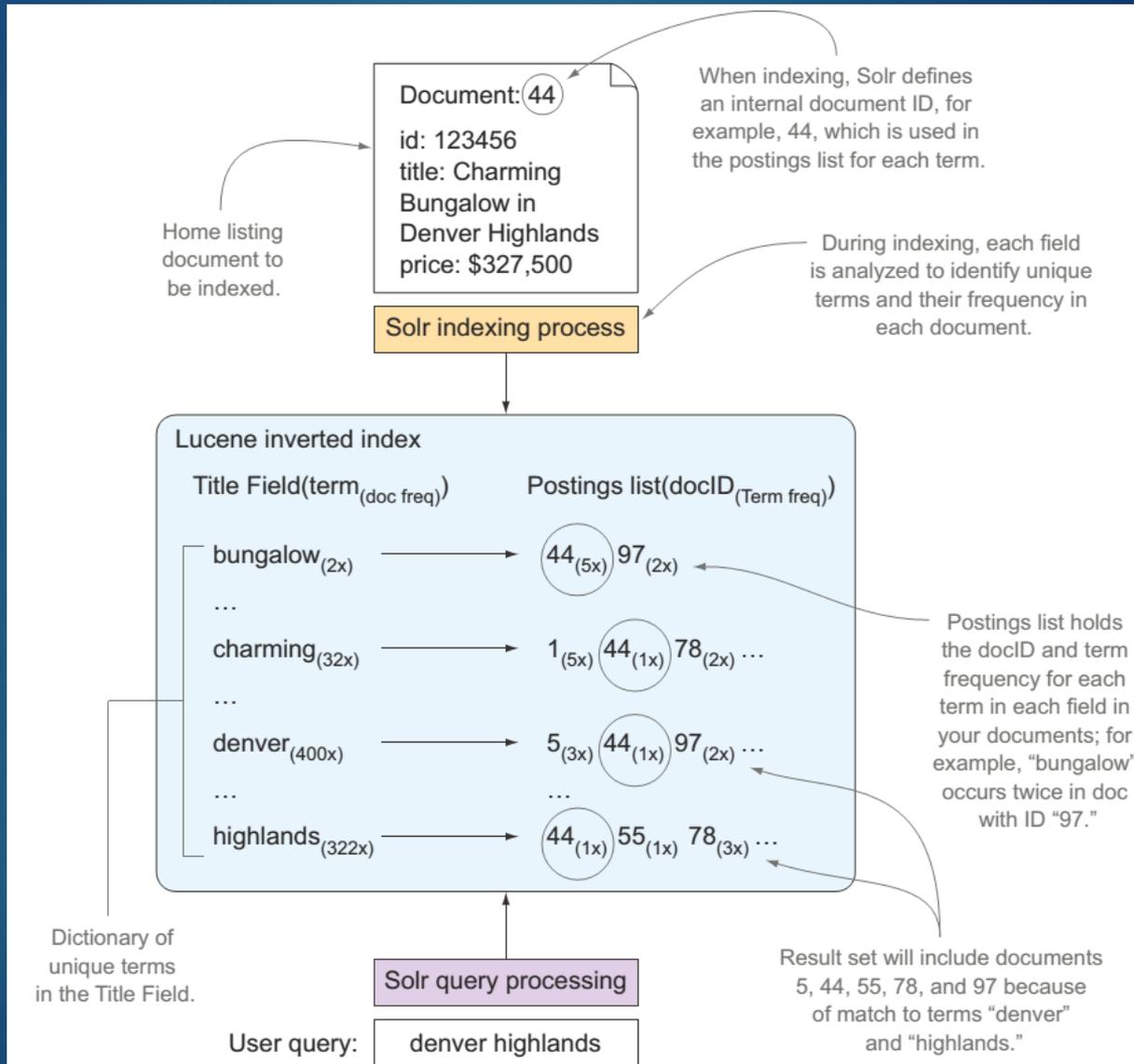
Solr

Introduzione

- ▶ **Apache Solr** è un motore di ricerca standalone scritto in java che utilizza il progetto Apache Lucene come backend.
- ▶ <http://lucene.apache.org/solr/>
- ▶ Solr è realizzato sotto forma di servlet e per funzionare necessita di un server per applicazioni web come Tomcat o Jetty
- ▶ Utilizzato pesantemente da: LinkedIn, Twitter, Netflix, Cnet, Digg



Indice Invertito



Solr Admin

The find all documents query in Solr is *.*

Search results from executing the find all documents query.

Apache Solr

Dashboard
Logging
Core Admin
Java Properties
Thread Dump
collection1
Overview
Analysis
Config
Dataimport
Documents
Ping
Plugins / Stats
Query
Replication
Schema
Schema Browser

Request-Handler (qt)
/select

common

q
.

fq

sort

start, rows
0 10

fl

df

Raw Query Parameters
key1=val1&key2=val2

wt
xml

indent
 debugQuery

dismax
 edismax
 hl
 facet
 spatial
 spellcheck

Execute Query

http://localhost:8983/solr/collection1/select?q=*&wt=xml&indent=true

```
<?xml version="1.0" encoding="UTF-8"?>
<response>
  <lst name="responseHeader">
    <int name="status">0</int>
    <int name="QTime">0</int>
    <lst name="params">
      <str name="indent">true</str>
      <str name="q">*.*</str>
      <str name="_">1376584164510</str>
      <str name="wt">xml</str>
    </lst>
  </lst>
  <result name="response" numFound="32" start="0">
    <doc>
      <str name="id">GB18030TEST</str>
      <str name="name">Test with some GB18030 encoded characters</str>
      <arr name="features">
        <str>No accents here</str>
        <str>这是一个功能</str>
        <str>This is a feature (translated)</str>
        <str>这份文件是很有光泽</str>
        <str>This document is very shiny (translated)</str>
      </arr>
      <float name="price">0.0</float>
      <str name="price_a">0,USD</str>
      <bool name="inStock">true</bool>
      <long name="_version_">1443452384713900032</long></doc>
    <doc>
      <str name="id">SP2514N</str>
      <str name="name">Samsung SpinPoint P120 SP2514N - hard drive - 250 GB - ATA-133</str>
      <str name="manu">Samsung Electronics Co. Ltd.</str>
      <str name="manu_id_s">samsung</str>
      <arr name="cat">
        <str>electronics</str>
        <str>hard drive</str>
      </arr>
      <arr name="features">
```

Open the core-specific tools for collection1 to find the link to the query form.

Indicizzazione e ricerca

▶ **Indicizzazione dei documenti**

- ▶ HTTP post di documenti XML
- ▶ DataImportHandler per basi di dati
- ▶ SolrCell per documenti complessi (PDF o Office)

▶ Ricerca

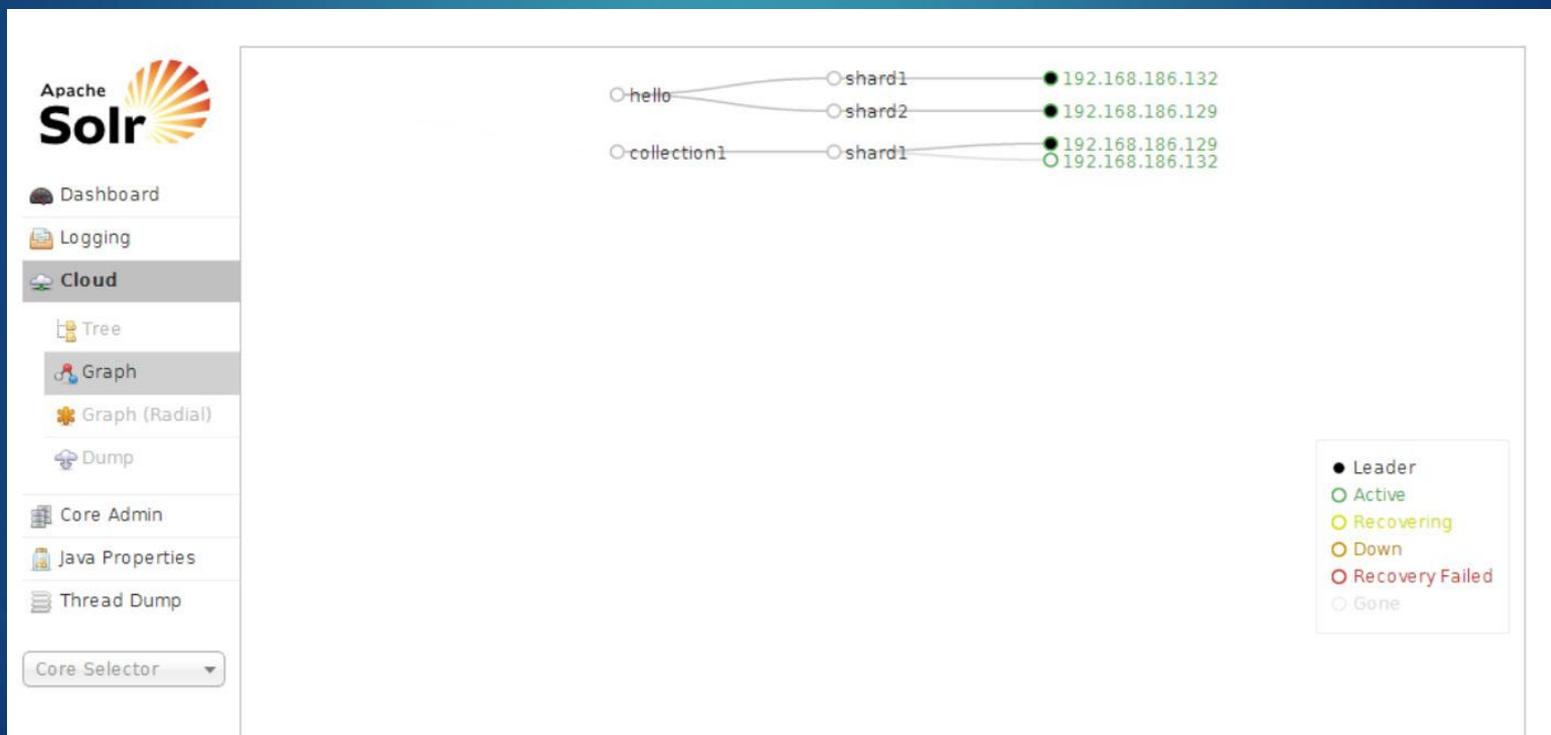
- ▶ Tramite URL:

- ▶ Es.

[http://localhost:8983/solr/select?indent=on&version=2.2&q=%3A&start=0&rows=10&fl=%2Cscore&qf=standard&wt=standard](http://localhost:8983/solr/select?indent=on&version=2.2&q=%3A*&start=0&rows=10&fl=%2Cscore&qf=standard&wt=standard)*

SolrCloud

- ▶ Memorizzazione indice su HDFS
- ▶ Segmentazione e replicazione core su più nodi
- ▶ Gestione automatica sharding



The screenshot displays the Apache Solr Cloud interface. On the left is a navigation sidebar with the following items: Dashboard, Logging, Cloud (selected), Tree, Graph (selected), Graph (Radial), Dump, Core Admin, Java Properties, and Thread Dump. At the bottom of the sidebar is a 'Core Selector' dropdown menu.

The main area shows a graph view of the cluster configuration. It features two collections: 'hello' and 'collection1'. The 'hello' collection has two shards: 'shard1' and 'shard2'. The 'collection1' collection has one shard: 'shard1'. Each shard is connected to one or more nodes, represented by colored circles with IP addresses. A legend in the bottom right corner defines the node colors: ● Leader (black), ○ Active (green), ○ Recovering (yellow), ○ Down (orange), ○ Recovery Failed (red), and ○ Gone (grey).

The graph shows the following connections:

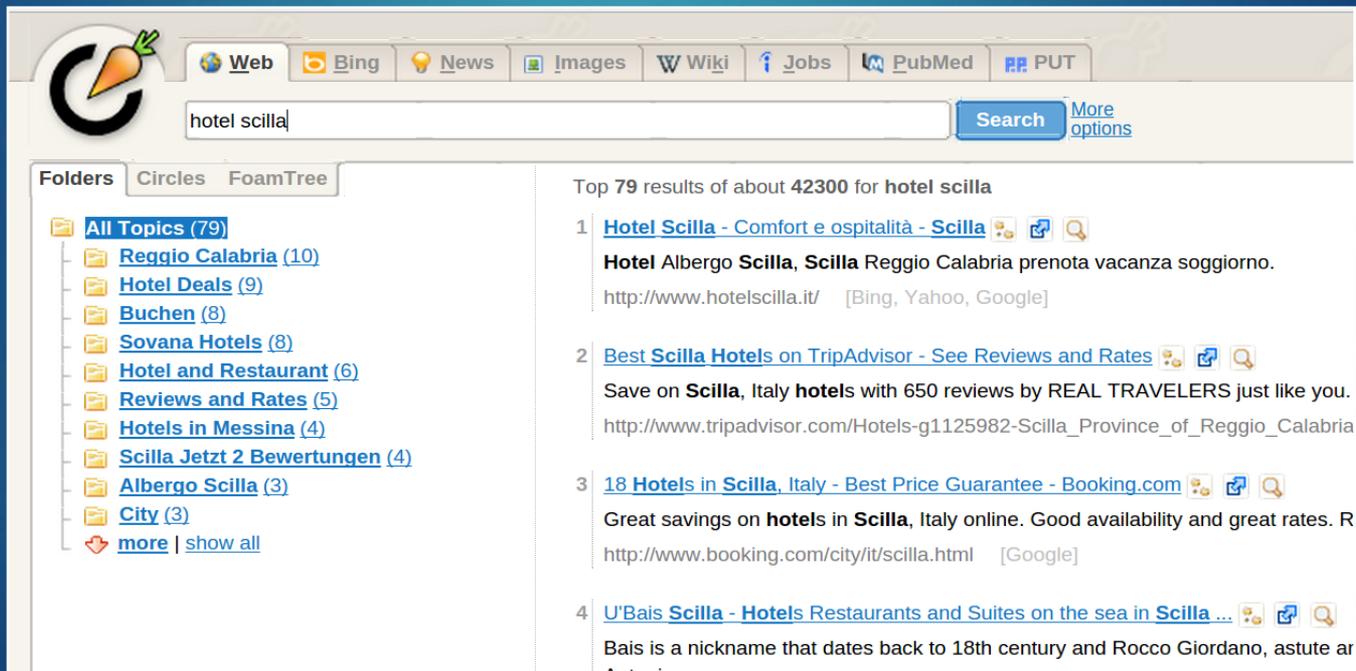
- 'hello' shard1 is connected to 192.168.186.132 (Leader).
- 'hello' shard2 is connected to 192.168.186.129 (Leader).
- 'collection1' shard1 is connected to 192.168.186.129 (Active) and 192.168.186.132 (Active).



Carrot2

Introduzione

- ▶ **Carrot2** è una libreria con la quale è possibile clusterizzare risultati provenienti da motori di ricerca
- ▶ Nessuna tassonomia o contenuti preclassificati
- ▶ Possibilità di interfacciamento con Solr o motori di ricerca già esistenti come Google o Bing



The screenshot displays the Carrot2 search interface. At the top left is the Carrot2 logo, a stylized carrot. Below it is a navigation bar with tabs for Web, Bing, News, Images, Wiki, Jobs, PubMed, and PUT. A search bar contains the text "hotel scilla" and a "Search" button with a "More options" link. Below the search bar are tabs for "Folders", "Circles", and "FoamTree". The "Folders" tab is active, showing a tree structure of folders with item counts: All Topics (79), Reggio Calabria (10), Hotel Deals (9), Buchen (8), Sovana Hotels (8), Hotel and Restaurant (6), Reviews and Rates (5), Hotels in Messina (4), Scilla Jetzt 2 Bewertungen (4), Albergo Scilla (3), and City (3). A "more" link and "show all" link are also visible. The main content area shows "Top 79 results of about 42300 for hotel scilla". The first four results are listed:

- 1 [Hotel Scilla - Comfort e ospitalità - Scilla](#) [Icons]
Hotel Albergo Scilla, Scilla Reggio Calabria prenota vacanza soggiorno.
<http://www.hotelscilla.it/> [Bing, Yahoo, Google]
- 2 [Best Scilla Hotels on TripAdvisor - See Reviews and Rates](#) [Icons]
Save on **Scilla, Italy hotels** with 650 reviews by REAL TRAVELERS just like you.
http://www.tripadvisor.com/Hotels-g1125982-Scilla_Province_of_Reggio_Calabria
- 3 [18 Hotels in Scilla, Italy - Best Price Guarantee - Booking.com](#) [Icons]
Great savings on **hotels in Scilla, Italy** online. Good availability and great rates. R
<http://www.booking.com/city/it/scilla.html> [Google]
- 4 [U'Bais Scilla - Hotels Restaurants and Suites on the sea in Scilla ...](#) [Icons]
Bais is a nickname that dates back to 18th century and Rocco Giordano, astute ar
Antonia

Algoritmi di Clustering

► Algoritmi integrati in Carrot2

- Lingo
- STC
- K-Means

Caratteristica	Lingo	STC	k-means
Diversità dei cluster	Alta. Vengono evidenziati molti piccoli clusters	Bassa. Pochi cluster evidenziati	Bassa. Pochi cluster evidenziati
Etichette dei Clusters	Lunghe, dunque molto descrittive.	Corte, ma ancora appropriate	Una sola parola ma non sempre descrittiva di tutti i documenti nel cluster.
Scalabilità	Bassa. Per un numero superiore a 1000 documenti, l'algoritmo Lingo potrebbe impiegare molto tempo e molta memoria.	Alta	Bassa, basata su strutture dati simili a Lingo.

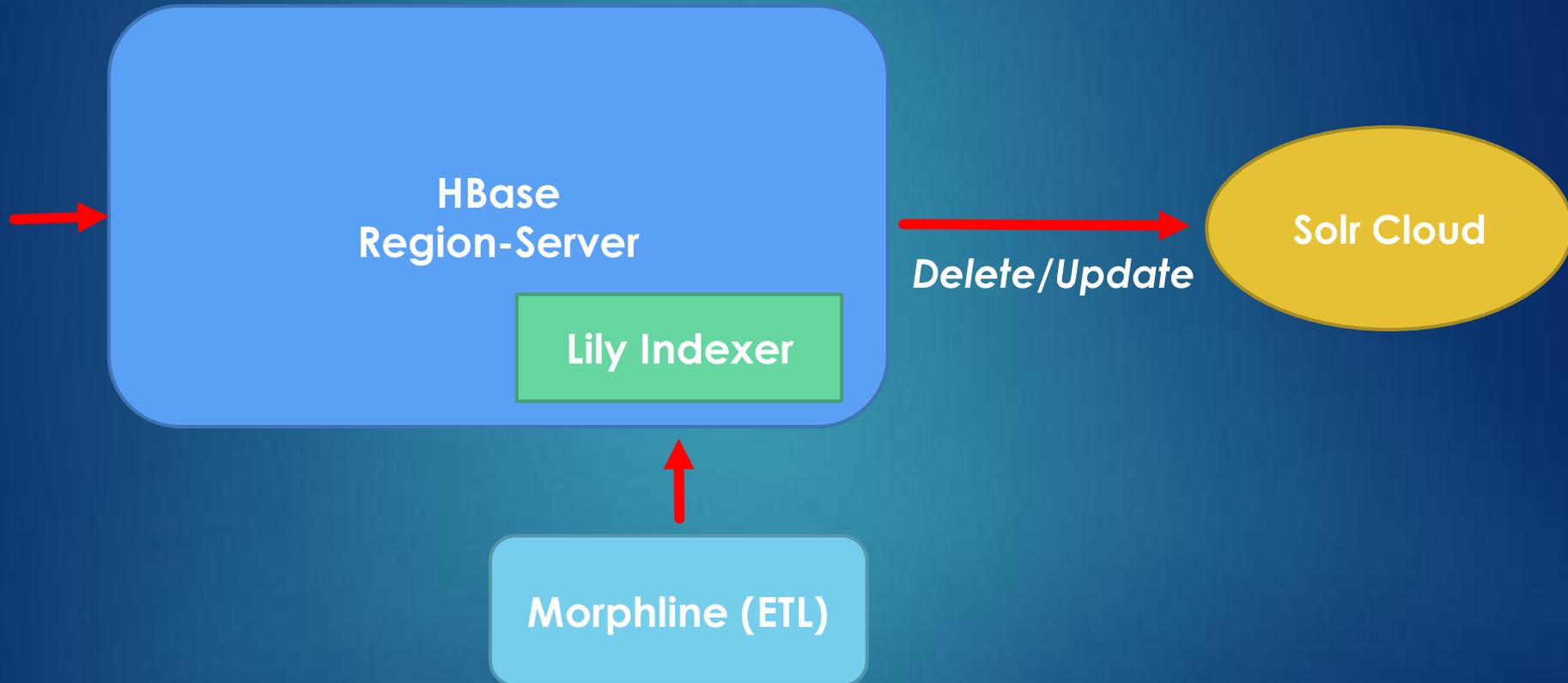


Indicizzazione Lily / Map-Reduce

Introduzione

- ▶ **Lily** è un sistema di indicizzazione NRT (Near Real Time) per HBase
- ▶ Rappresenta una valida soluzione per ottenere una ricerca Full-Text efficiente dei dati memorizzati in HBase
- ▶ Simula il comportamento di un Hbase Region Server, intercettando le richieste di scrittura e cancellazione, delegandole a Solr Cloud (dopo aver applicato un ETL “al volo”)

L'architettura di Lily



Un esempio di morphline per l'ETL

```
{
  inputColumn : "cfAnagrafica:sorgenti"
  outputField : "sorgenti"
  type : "string"
  source : value
}
{
  inputColumn : "cfAttributi:servizio_*"
  outputField : "servizi"
  type : "string"
  source : value
}
]
}
}

{
  addValues { location : "@{latitudine},@{longitudine}" }
}
```

Indicizzazione con Map Reduce

Nei casi in cui è necessario indicizzare dati in maniera statica (per esempio caricamento iniziale del DW) è conveniente utilizzare il Map Reduce

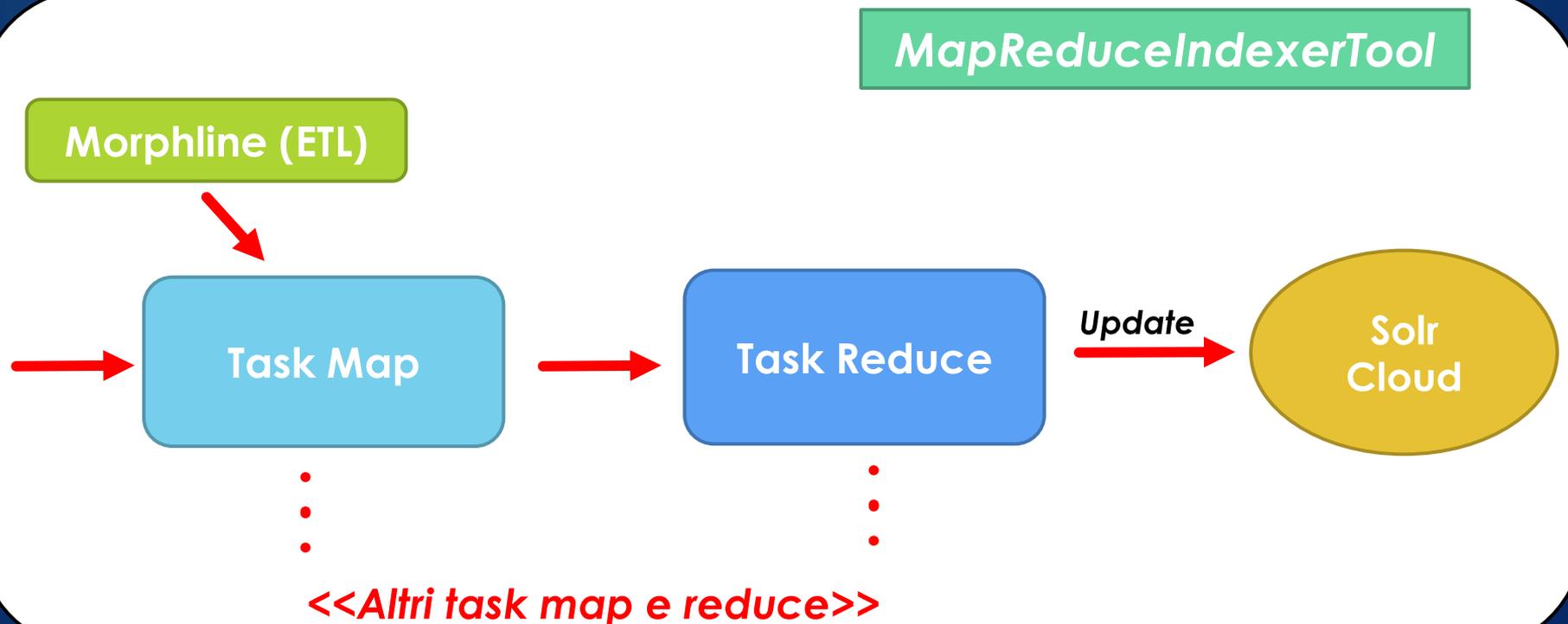
In questo caso:

- ▶ la Map produce una coppia termine univoco e ID del documento in cui il termine è stato trovato
- ▶ mentre nella fase di Reduce ci si ritrova per ogni termine la lista dei documenti in cui esso occorre

Perciò la Reduce non dovrà fare altro che contare per ogni termine le relative frequenze nei documenti e costruire “l’indice invertito”

Architettura di Map Reduce Indexing

Come per Lily viene utilizzato un morphline per eseguire dell'ETL sui dati prima del loro caricamento su Solr





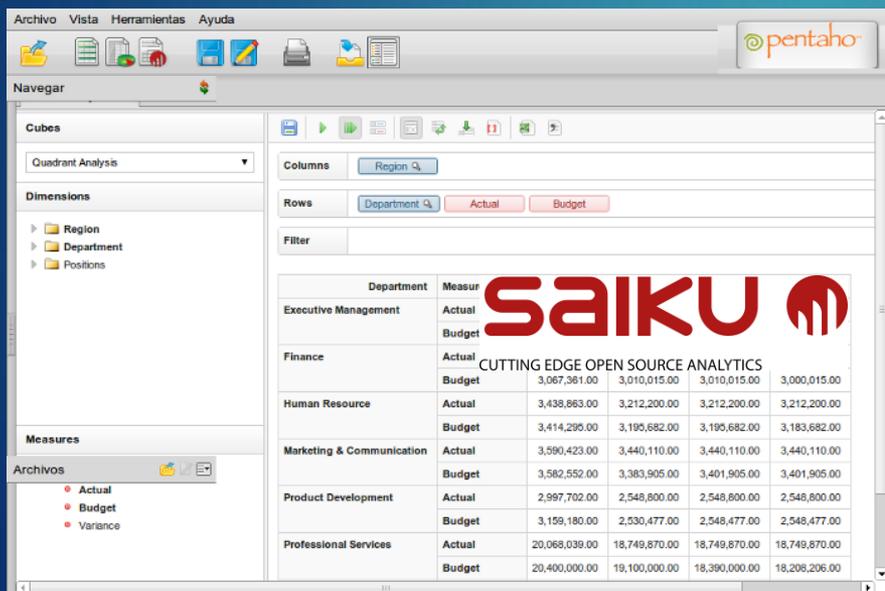
Architettura per Analisi Batch

Data Analysis Tools



Mondrian

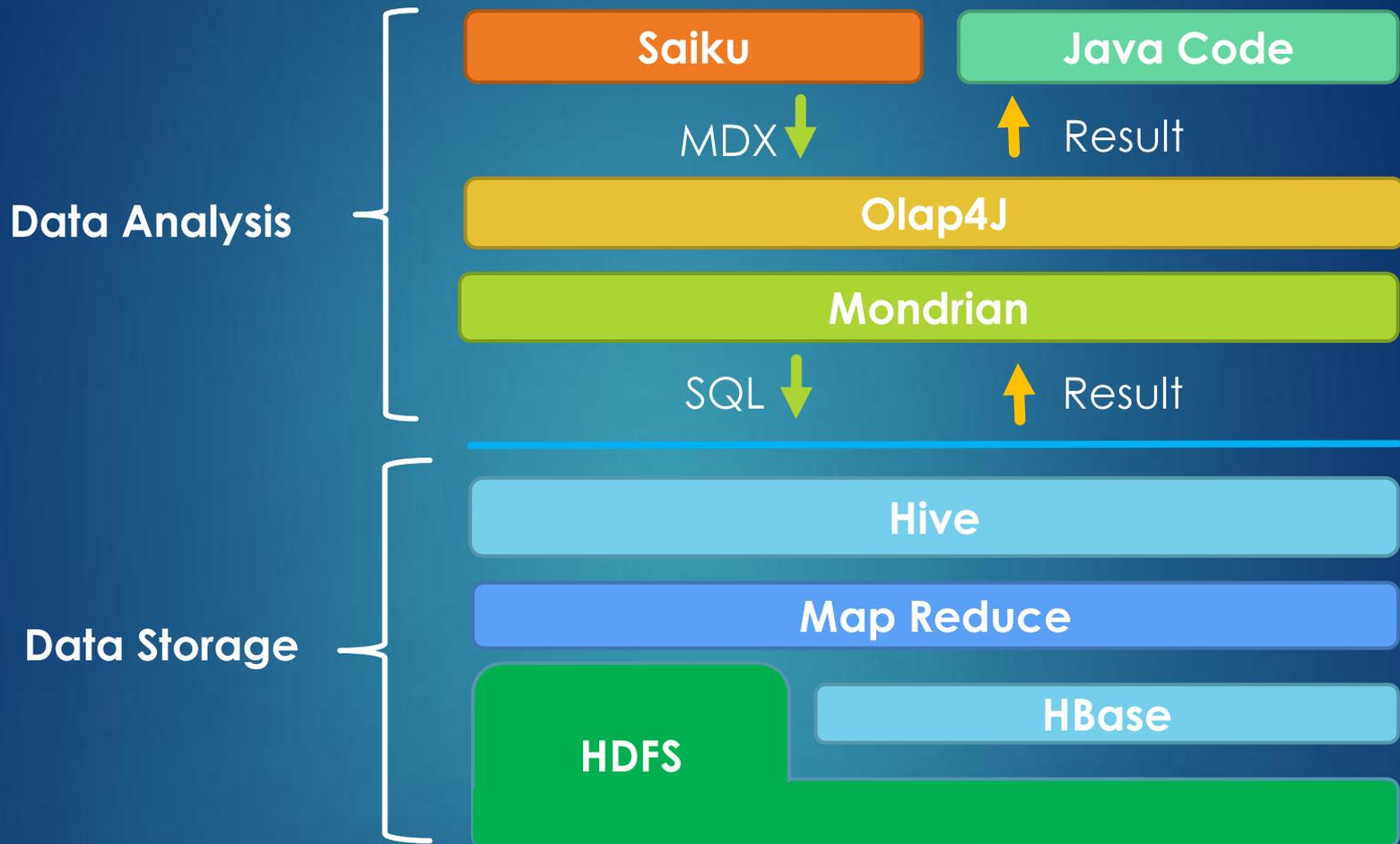
Mondrian è l'OLAP server fornito da Pentaho. Si basa su tecnologia ROLAP, consente di tradurre query MDX in SQL, basandosi su un modello multidimensionale definito dagli utenti.



Saiku

Saiku è un plugin che offre molte possibilità di visualizzazione dei dati in Pentaho. Consente la selezione dei dati tramite drag and drop

Architettura del sistema (generazione reports)



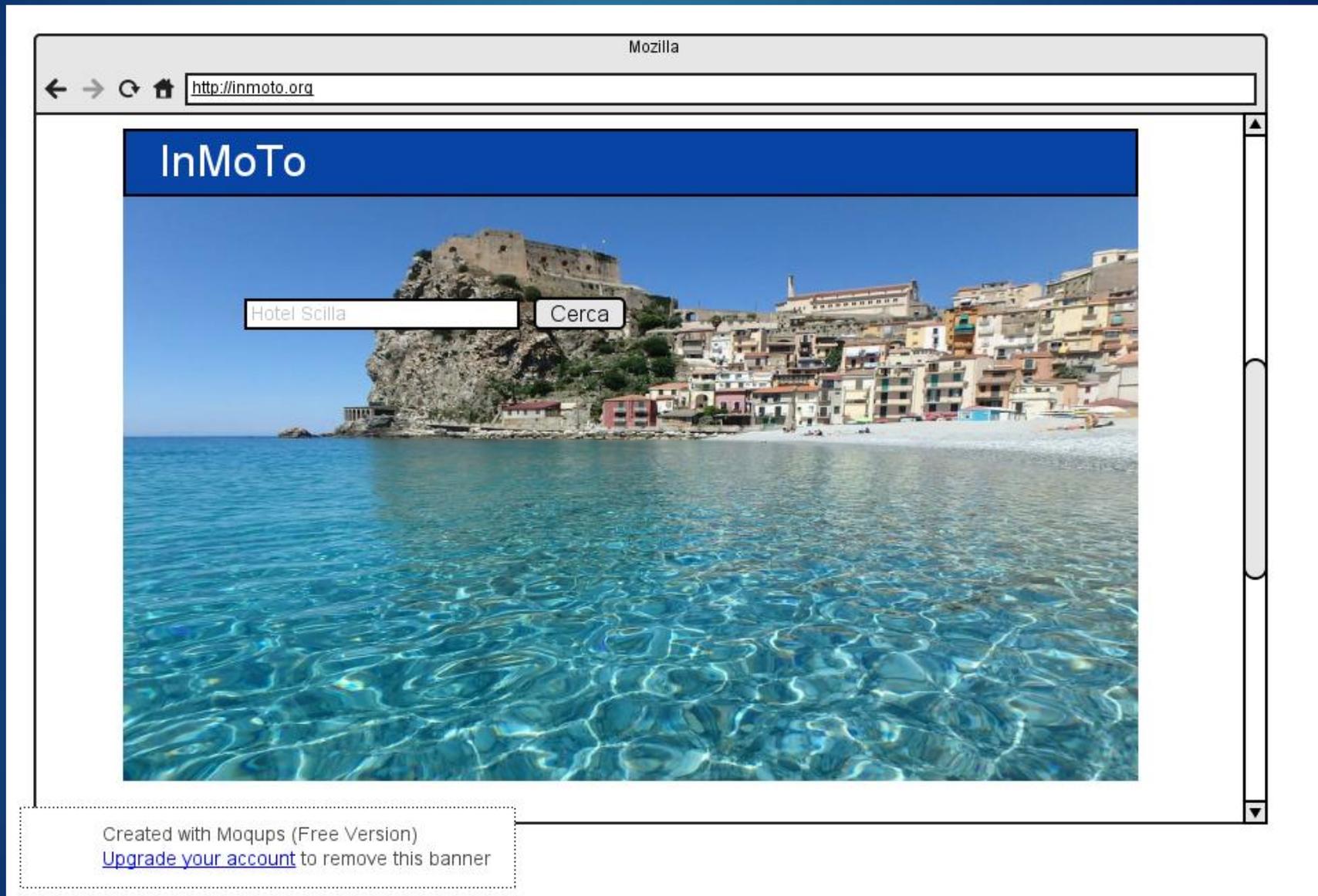
L'architettura proposta è composta da 3 moduli:

1. **Mondrian come OLAP Server**
2. **Hive come esecutore di query su Map Reduce**
3. **HBase o HDFS come Data Storage**



Architettura per Ricerca Full Text

Caso d'uso: Ricerca



Caso d'uso: Visualizzazione Risultati

Mozilla

← → ↻ 🏠

InMoTo

Stelle

- 1 (80)
- 2 (23)
- 3 (112)
- 4 (4)

Prezzo

- < 50€ (80)
- Fra 50€ e 80€ (23)
- Fra 80€ e 100€ (112)
- 100€ (4)

Carta Credito



Clusters

- ▼ Primi Piatti (80)
 - ▶ Buoni (49)
 - ▶ Pessimi (23)
 - ▶ Crudi (8)
- ▶ Da asporto (64)
- ▶ Pessimi Dolci (12)
- ▶ Carne buona (31)

Hotel Sette Notti

 ★★★★★
 www.settenotti.it
Lorem ipsum dolor sit amet, consectetur adipiscing elit. Nulla quam velit, vulputate eu pharetra nec, mattis ac neque. Duis vulputate commodo lectus, ac blandit elit

Hotel Scilla Day

 ★★★★★
 www.scilladay.it
Lorem ipsum dolor sit amet, consectetur adipiscing elit. Nulla quam velit, vulputate eu pharetra nec, mattis ac neque. Duis vulputate commodo lectus, ac blandit elit

Hotel Papero

 ★★★★★
 www.papero.com
Lorem ipsum dolor sit amet, consectetur adipiscing elit. Nulla quam velit, vulputate eu pharetra nec, mattis ac neque. Duis vulputate commodo lectus, ac blandit elit

B&B Tallori

 ★★★★★
 www.tallori.org
Lorem ipsum dolor sit amet, consectetur adipiscing elit. Nulla quam velit, vulputate eu pharetra nec, mattis ac neque. Duis vulputate commodo lectus, ac blandit elit

Created with Moqups (Free Version)
[Upgrade your account](#) to remove this banner

Mozilla

← → ↻ 🏠

InMoTo

Hotel Sette Notti

★★★★☆



www.settenotti.it
Lorem ipsum dolor sit amet, consectetur adipiscing elit. Nulla quam velit, vulputate eu pharetra nec, mattis ac neque. Duis vulputate commodo lectus, ac blandit elit tincidunt id. Sed rhoncus, tortor sed eleifend tristique, tortor mauris molestie elit, et lacinia ipsum quam nec dui. Quisque nec mauris sit amet elit iaculis pretium sit amet quis magna. Aenean velit odio, elementum in tempus ut, vehicula eu diam. Pellentesque

Info

Accetta Carta	No
Prezzo Medio	80 €
Parcheggio	Si
TV Satellitare	No
Ha frigoBar	Si
Wi-Fi	Si

Recensioni

 www.papero.com
Lorem ipsum dolor sit amet, consectetur adipiscing elit. Nulla quam velit, vulputate eu pharetra nec, mattis ac neque. Duis vulputate commodo lectus, ac blandit elit

 www.tallori.org
Lorem ipsum dolor sit amet, consectetur adipiscing elit. Nulla quam velit, vulputate eu pharetra nec, mattis ac neque. Duis vulputate commodo lectus, ac blandit elit



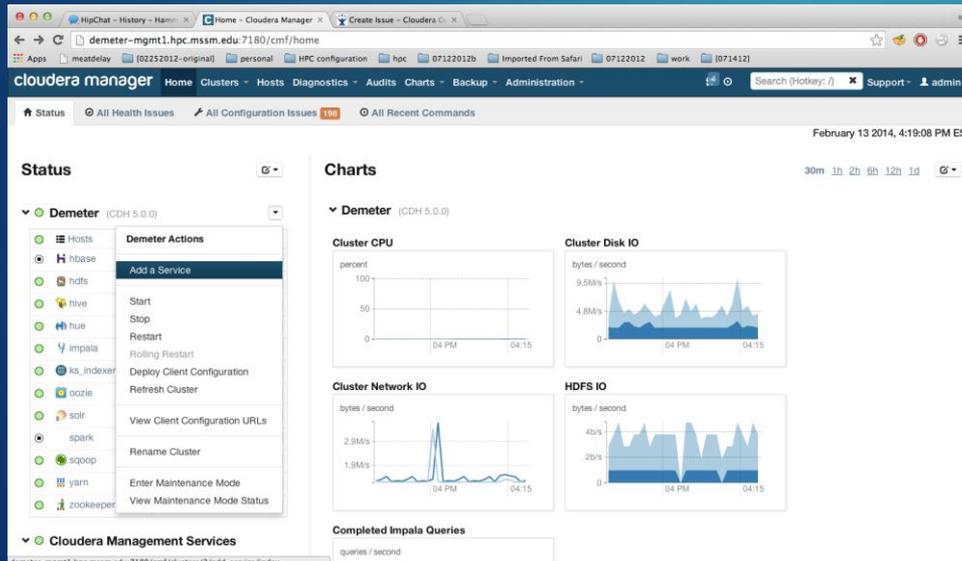
La suite Cloudera

La Suite



Cloudera è una suite completa di tutti i componenti piu' utili all'interazione con l'ecosistema Hadoop. Viene chiamata brevemente **CDH** (Cloudera Distribution for Apache Hadoop) ed include un manager che permette tramite browser, l'installazione e la gestione dei vari servizi sui nodi del cluster.

Hosts	1
hbase1	
hdfs1	
hive1	
hue1	
impala1	
mapreduce1	
oozie1	
sqoop1	
yarn1	
zookeeper1	



- HDFS
- HBase
- Hadoop (Map-Reduce -Yarn)
- Hive
- Impala
- Sqoop
- Hue
- Zookeeper
-

Requisiti

L'installazione di Cloudera richiede l'uso delle più comuni distribuzioni Linux (Centos, Ubuntu, etc..)

Per una installazione minimale è necessario avere tre macchine con la seguente configurazione:

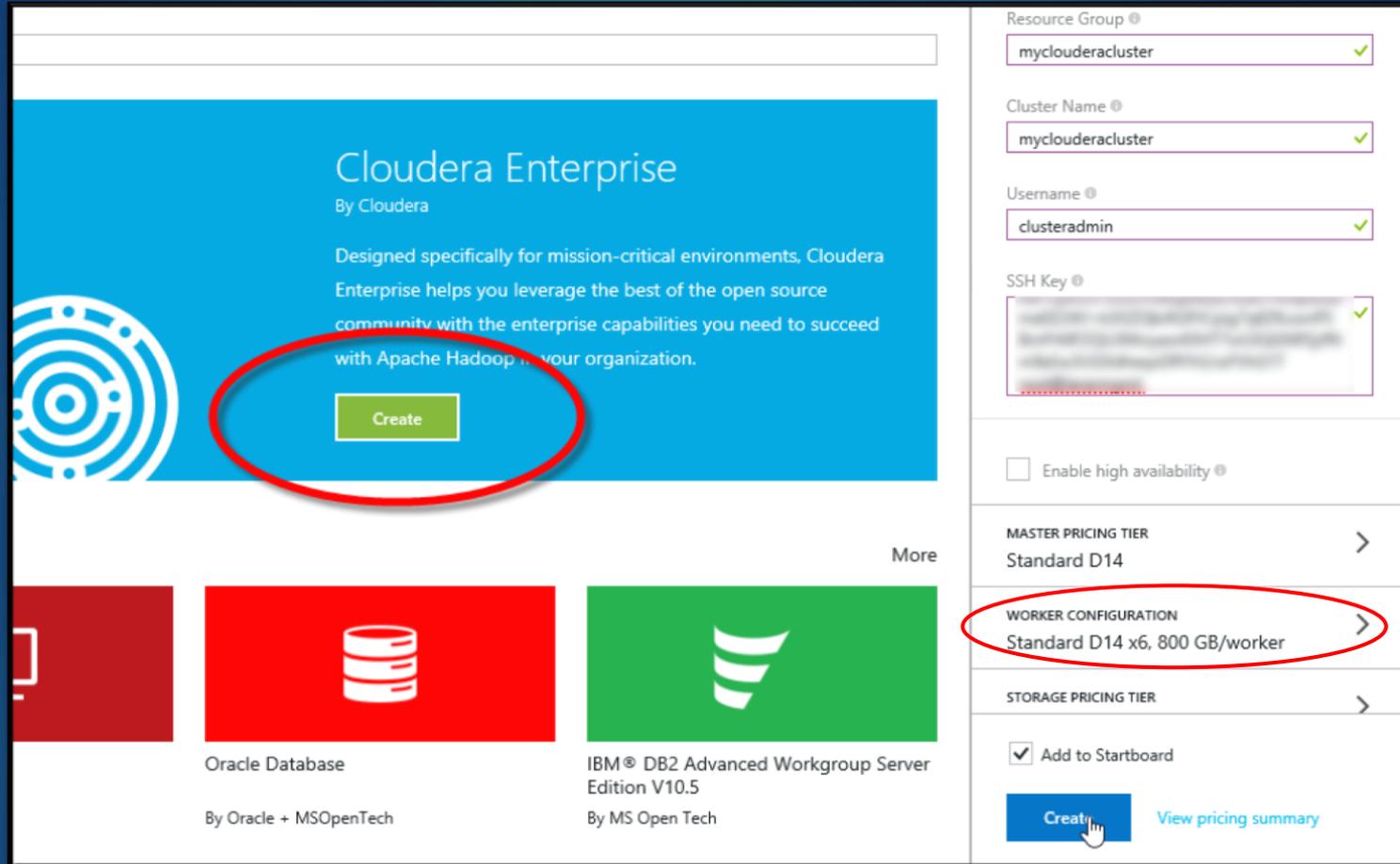
- ▶ 8 GB RAM sul nodo master e 4GB sui due nodi slave
- ▶ CPU con due core
- ▶ 100 GB di hardisk a macchina (N.B. collo di bottiglia del sistema)

I requisiti raccomandati prevedono almeno 10 nodi con la seguente configurazione:

- ▶ 32 GB RAM sul nodo master e 16GB sui nodi slave
- ▶ CPU con 8 core sul master e almeno 4 sui nodi slave
- ▶ 3X500GB di hardisk SSD "dedicato"
- ▶ Rete Gigabit

L'installazione di Cloudera è anche supportata "nativamente" dalle soluzioni di cloud più note (per es. Microsoft Azure)

Cloudera su Windows Azure



Cloudera Enterprise
By Cloudera

Designed specifically for mission-critical environments, Cloudera Enterprise helps you leverage the best of the open source community with the enterprise capabilities you need to succeed with Apache Hadoop in your organization.

Create

Oracle Database
By Oracle + MSOpenTech

IBM® DB2 Advanced Workgroup Server Edition V10.5
By MS Open Tech

Resource Group 
myclouderacluster 

Cluster Name 
myclouderacluster 

Username 
clusteradmin 

SSH Key 
 

Enable high availability 

MASTER PRICING TIER 
Standard D14

WORKER CONFIGURATION 
Standard D14 x6, 800 GB/worker

STORAGE PRICING TIER 

Add to Startboard

Create  [View pricing summary](#)

ISTANZA	CORE	RAM	DIMENSIONI DEI DISCHI	PREZZO
D14	16	112 GB	800 GB	€1,9445/ora (~€1.447 al mese)



DEMO

Yelp su Solr

Yelp è un **social network** in cui le persone si scambiano **pareri, opinioni** e “**dritte**” sui posti migliori del luogo in cui abitano e di quelli in cui vanno per lavoro, viaggio o altri motivi.



```
<field name="id" type="string" indexed="true" stored="true" required="true" />
<field name="phone" type="string" indexed="true" stored="true" />
<field name="review_count" type="tint" indexed="true" stored="true" />
<field name="is_closed" type="boolean" indexed="true" stored="true" />
<field name="is_claimed" type="boolean" indexed="true" stored="true" />
<field name="url" type="string" indexed="true" stored="true" />
<field name="name" type="text" indexed="true" stored="true" />
<field name="mobile_url" type="string" indexed="true" stored="true" />
<field name="categories" type="string" indexed="true" stored="true"/>
<field name="rating" type="tfloat" indexed="true" stored="true" />
<field name="image_url" type="string" indexed="true" stored="true" />
<field name="display_phone" type="text" indexed="true" stored="true" />
<field name="distance" type="double" indexed="true" stored="true" />
<field name="snippet_text" type="text" indexed="true" stored="true" />
<field name="snippet_image_url" type="text" indexed="true" stored="true" />
<field name="location" type="text" indexed="true" stored="true" />
<field name="review_id" type="string" indexed="true" stored="true" multiValued="true" />
<field name="review_rating" type="tfloat" indexed="true" stored="true" multiValued="true" />
<field name="review_excerpt" type="text" indexed="true" stored="true" multiValued="true" />
<field name="review_time_created" type="long" indexed="true" stored="true" multiValued="true" />
<field name="user_id" type="string" indexed="true" stored="true" multiValued="true"/>
<field name="user_image_url" type="string" indexed="true" stored="true" multiValued="true"/>
<field name="user_name" type="string" indexed="true" stored="true" multiValued="true"/>
```

Problemi legati all'estrazione dei dati (imitidelle API):

- 25.000 richieste al giorno
- 1.000 risultati ad interrogazione
- 1 recensione per attività

```
http://192.168.0.16:8983/solr/yelp_shard1_replica1/clustering?q=text:ristorante+AND+text:milano&rows=100
```

Google Book n-grams

1/8



▶ N-grams

- ▶ Un **n-gramma** è una sottosequenza di n elementi di una data sequenza.
- ▶ In questo caso gli elementi sono parole estratte da Google Book

▶ Dataset

- ▶ File compressi plain text tab-separated
 - ▶ Ngram: l'attuale n-gramma
 - ▶ Year: anno di aggregazione
 - ▶ Occurrences: Numero di volte che l'n-gramma appare nell'anno
 - ▶ Book: numero di libri in cui appare l'n-gramma

Raggruppamento termini per decade con ratio

- ▶

```
CREATE TABLE by_decade (  
  gram string,  
  decade int,  
  ratio double  
);
```
- ▶

```
INSERT OVERWRITE TABLE by_decade  
SELECT  
  a.gram,  
  b.decade,  
  sum(a.occurrences) / b.total  
FROM  
  normalized a  
JOIN (  
  SELECT  
    substr(year, 0, 3) as decade,  
    sum(occurrences) as total  
  FROM  
    normalized  
  GROUP BY  
    substr(year, 0, 3)  
) b  
ON  
  substr(a.year, 0, 3) = b.decade  
GROUP BY  
  a.gram,  
  b.decade,  
  b.total;
```

Trend dei termini per decade

1/8



```
SELECT
  a.gram as gram,
  a.decade as decade,
  a.ratio as ratio,
  a.ratio / b.ratio as increase
FROM
  by_decade a
JOIN
  by_decade b
ON
  a.gram = b.gram and
  a.decade - 1 = b.decade
WHERE
  a.ratio > 0.0001 and
  a.decade >= 190
SORT BY
  decade ASC,
  increase DESC;
```

