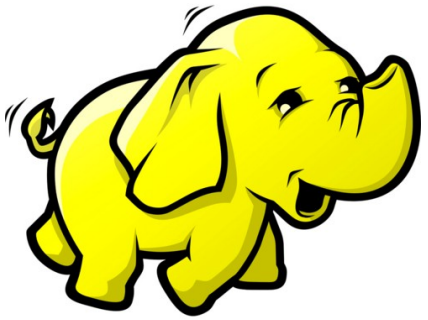


Introduzione ad Hadoop *

Hadoop.apache.org

** Presentazione tratta da: “Introducing Apache Hadoop: The Modern Data Operating System” di Dr. Amr Awadallah, fondatore e CTO di Cloudera*

Chi sono?

- Laureato triennale in Ingegneria Informatica all'Università di Modena e Reggio Emilia

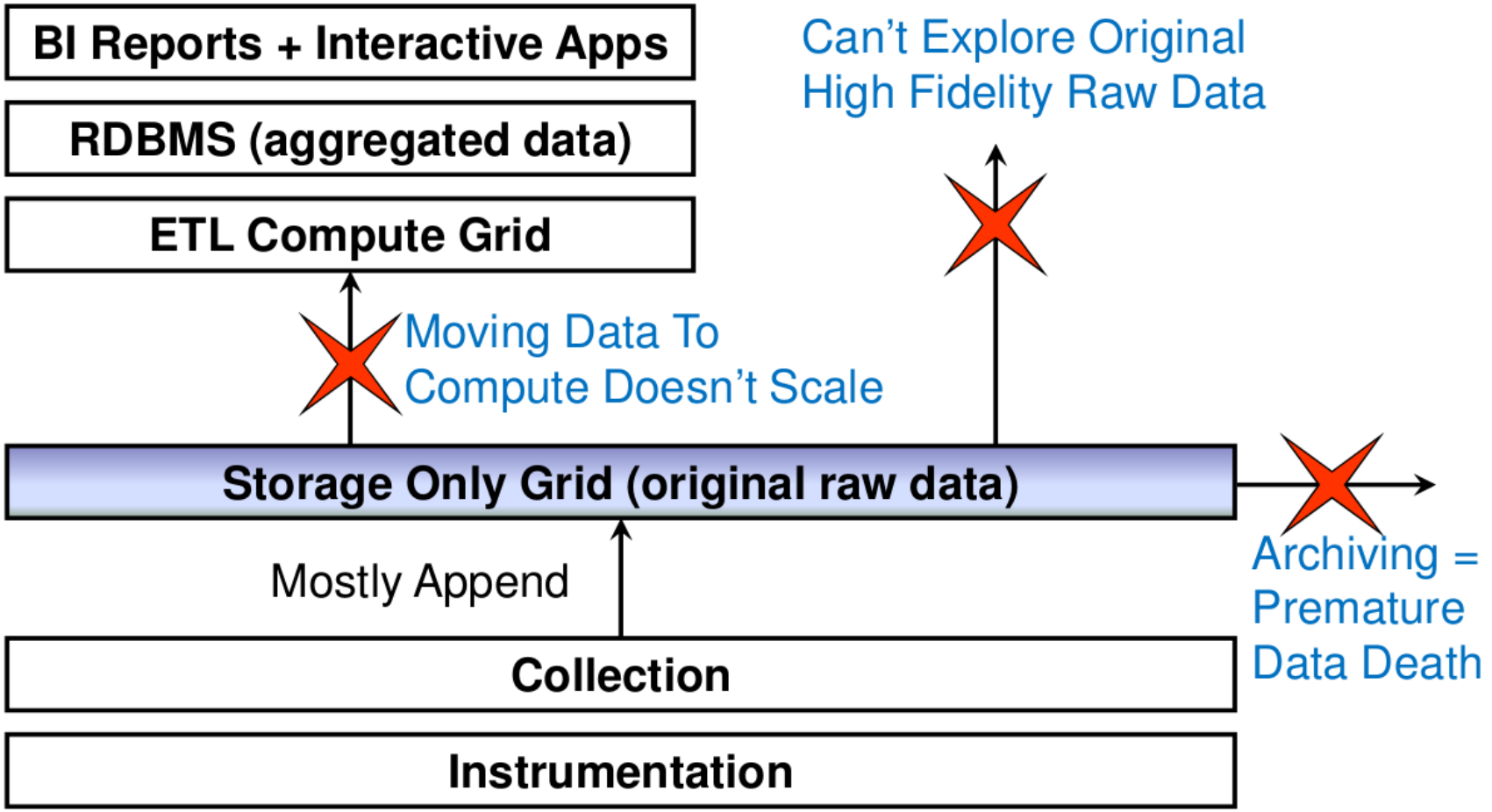
UNIVERSITÀ DEGLI STUDI
DI MODENA E REGGIO EMILIA

- Studente al corso magistrale al Politecnico di Milano

- Da ottobre 2012 collaboro con il Machine Learning Group dell'Université Libre De Bruxelles (Belgio) per la tesi di laurea magistrale

ULB UNIVERSITÉ LIBRE DE BRUXELLES

Limitazioni dell'architettura dati esistente



Che cos'è Apache Hadoop?

- È un sistema distribuito per il salvataggio e l'interrogazione dei dati, **scalabile** e capace di **gestire i guasti** (fault-tolerant)
- Progetto open source sotto licenza Apache
- Hadoop fornisce due principali sistemi:
  - **Hadoop Distributed File System** (HDFS): file system distribuito per salvare dati su un cluster di computer
  - **MapReduce**: paradigma di programmazione realizzato per offrire scalabilità e tolleranza ai guasti

Scalabilità

- Il programma scritto per Hadoop funziona a prescindere dalla dimensione del cluster
- Meglio avere più dati che un algoritmo più intelligente
  - A. Halevy et al, “The Unreasonable Effectiveness of Data”, IEEE Intelligent Systems, March 2009
- Possibilità di analizzare tutti i dati, invece che archiviare quelli più vecchi sul cassette.

Il vantaggio principale: Agilità/Flessibilità

Schema on Write (RDBMS)

- Lo schema deve essere creato prima che i dati vengano caricati
- Ogni dato da caricare deve essere trasformato nella struttura interna del DB
- Nuove colonne devono essere aggiunte esplicitamente prima che i nuovi dati per tali colonne siano caricate nel DB

Schema-on-Read (Hadoop)

- I dati sono semplicemente copiati nel file system, nessuna trasformazione è richiesta
- I dati delle colonne sono estratte durante la fase di lettura (*late binding*)
- I nuovi dati possono essere aggiunti ed estratti in qualsiasi momento

Lettura dati veloce
Standard

Pros

Caricamento dati veloce
Agilità/Flessibilità

Ad ognuno il suo lavoro

Database Relazionali

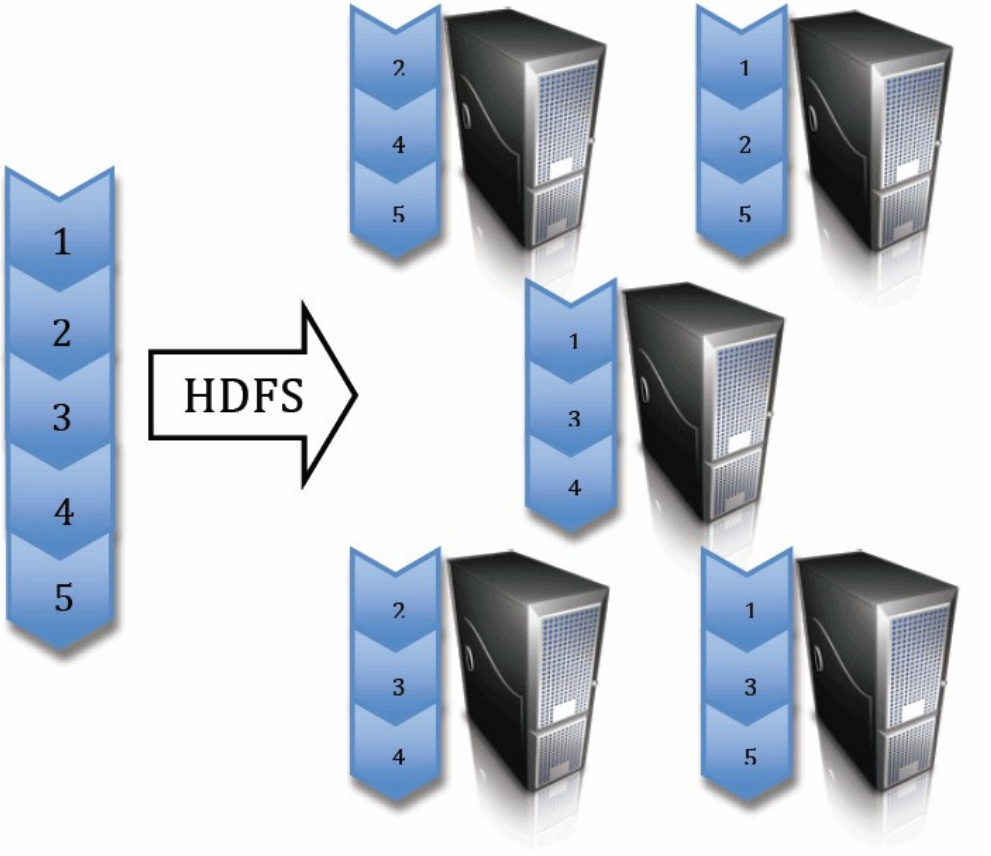
- Analisi OLAP (< 1 secondo)
- Transazioni complesse ACID
- Supporta al 100% SQL

Hadoop

- Dati strutturati e non (flessibilità)
- Scalabilità dei dati e della computazione
- Analisi complessa dei dati

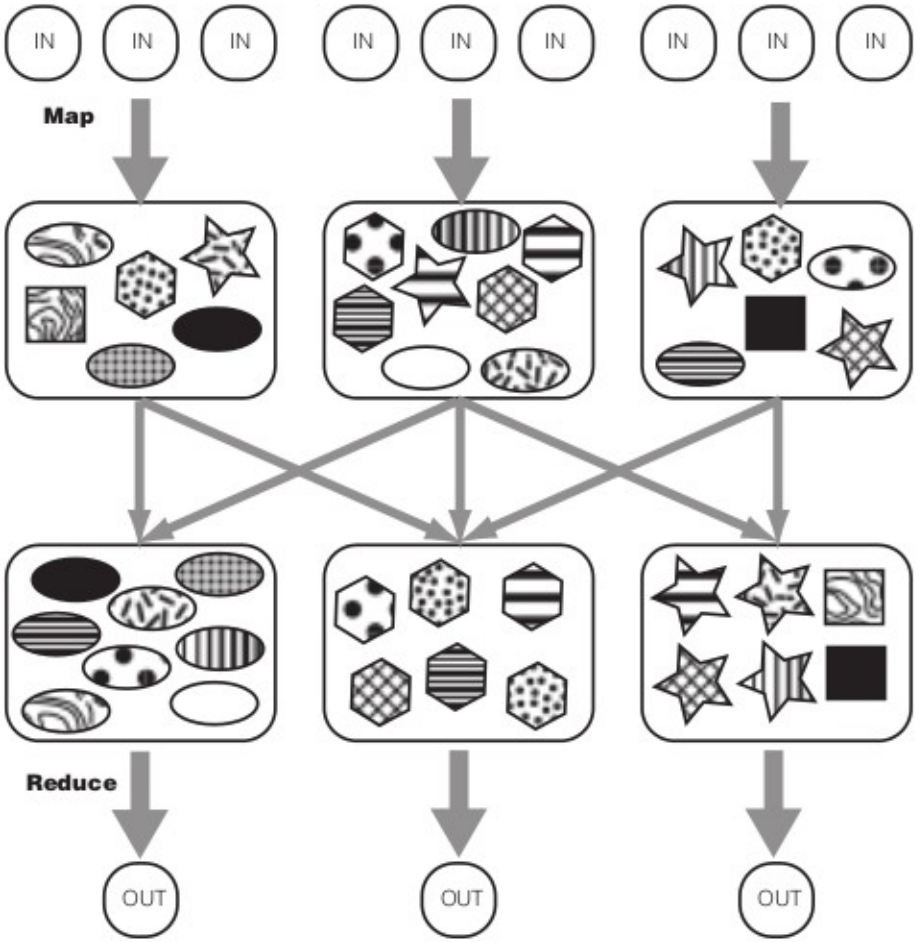
HDFS: Hadoop Distributed File System

- Ogni file è suddiviso in blocchi (default 64MB)
- Ogni blocco è replicato all'interno del cluster (default 3 copie)
  - Durability, Availability, Throughput
  - Le copie sono distribuite tra i server che tra i rack.
- Sistema ottimizzato per il throughput, per le operazione di Get/Delete/Append

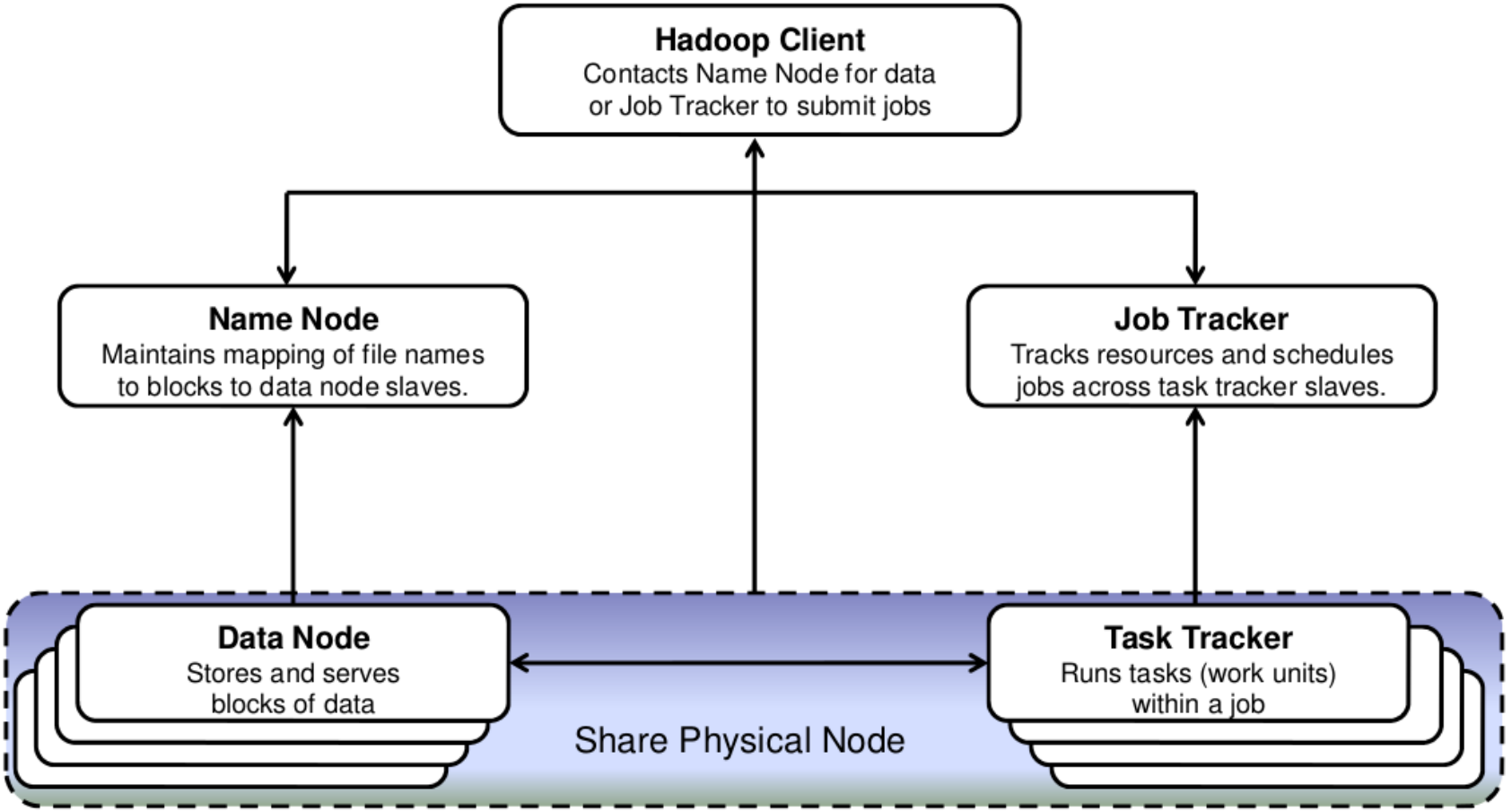


MapReduce

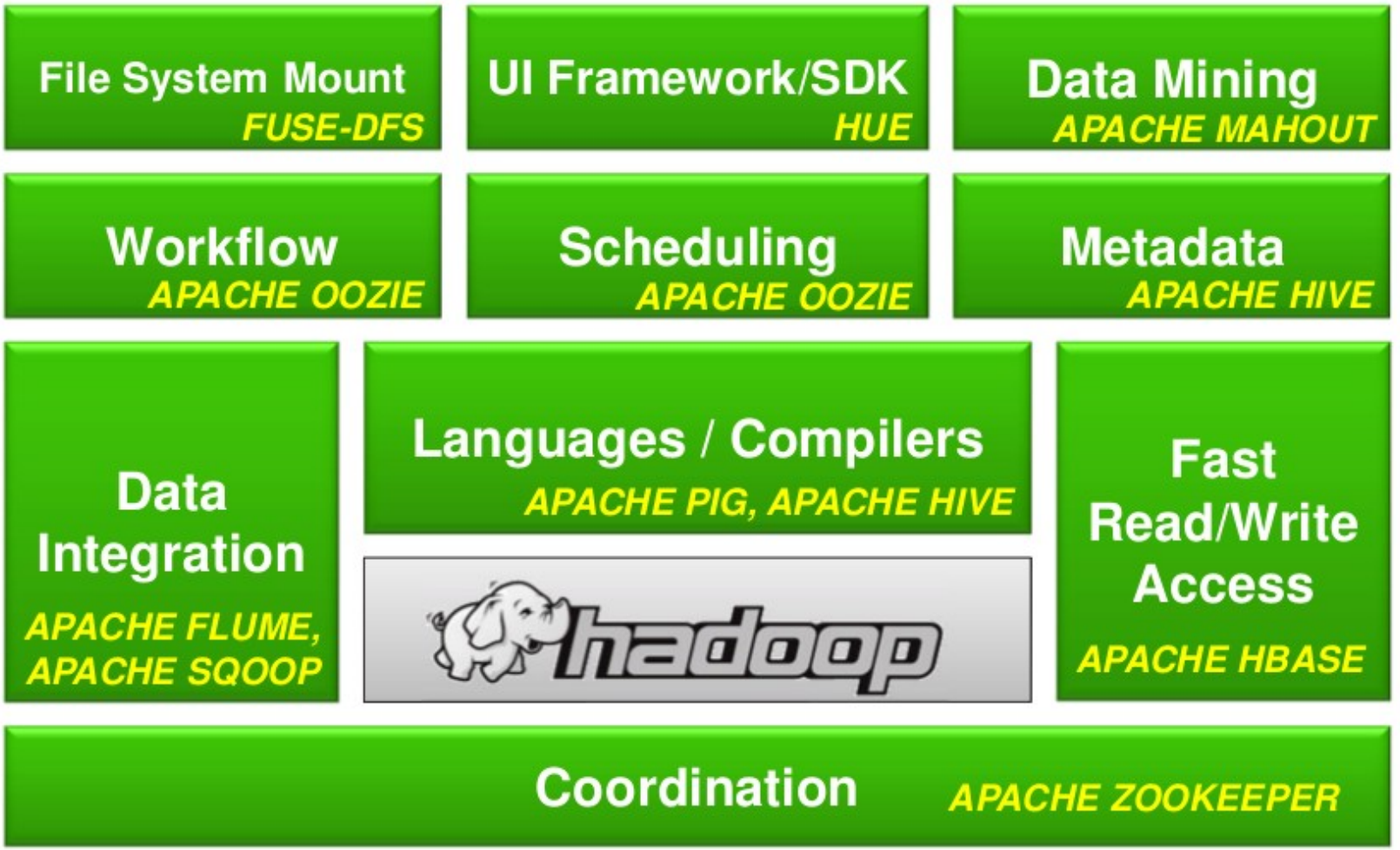
- Il formato dei dati è chiave-valore (<k,v>)
- Due fasi principali:
  - **Map**: i dati vengono mappati
  - **Reduce**: i dati vengono aggregati e processati
- Esempi:
  - Seleziona max valore
  - Conta parole (da un testo)



Architettura di Hadoop



Ecosistema Hadoop



Obiettivi raggiunti

- Implementato un algoritmo di **feature selection** (mRMR) su Hadoop
  - Deployment su Amazon EC2 e Amazon EMR

- Installazione e configurazione di Hadoop sul centro di calcolo universitario

Obiettivi raggiunti

- Implementato un algoritmo di **feature selection** (mRMR) su Hadoop
  - Deployment su Amazon EC2 e Amazon EMR
- Installazione e configurazione di Hadoop sul centro di calcolo universitario
- Sviluppare un modello teorico per il dimensionamento dell'infrastruttura informatica da utilizzare per l'esecuzione degli algoritmi
  - Ridurre di costi