

DATA LAKE - THE FOUR STAGES OF DATA MANAGEMENT

INGESTION & STORAGE



Real Time (data stream):

- Scalabilità automatica
- Monitoring in real time
- Modello di pricing "pay as you go"

Batch files (transactions, log-files)

- Gestione grossi volumi di dati in formato grezzo
- Governance del dato (regole di ingestion)
- Sicuro (accesso controllato ai dati), scalabile, affidabile (ridondato), storage a lungo termine
- Monitoring

PROCESSING



Manipolazione dei dati:

- Conversione dei dati in formati ottimizzati
- Estrazione di entità specifiche e parsing dei dati

Arricchimento dei dati:

- Incrocio di varie sorgenti dati
- Inferenza e machine learning

Automazione e Monitoraggio:

- Processi automatizzati e customizzabili per sorgente
- Monitoraggio ed sistema di alert automatico

ANALYSIS



Real time:

Per analizzare set di dati specifici velocemente e per restituire informazioni in tempo reale

Distribuito:

- Per analizzare grosse moli di dati in batch (sfrutta tecnologie Hadoop)
- Per realizzare modelli di machine learning per estrarre conoscenza dai dati

Esplorazione:

Per esplorare grosse moli di dati con un'ampia scelta di strumenti, come Hive, MySQL, Athena, Tableau.

INTEGRATION



API:

Per trasferire informazioni ad altre piattaforme / Apps

Data Visualization:

Per analizzare i dati e visualizzarli tramite reports

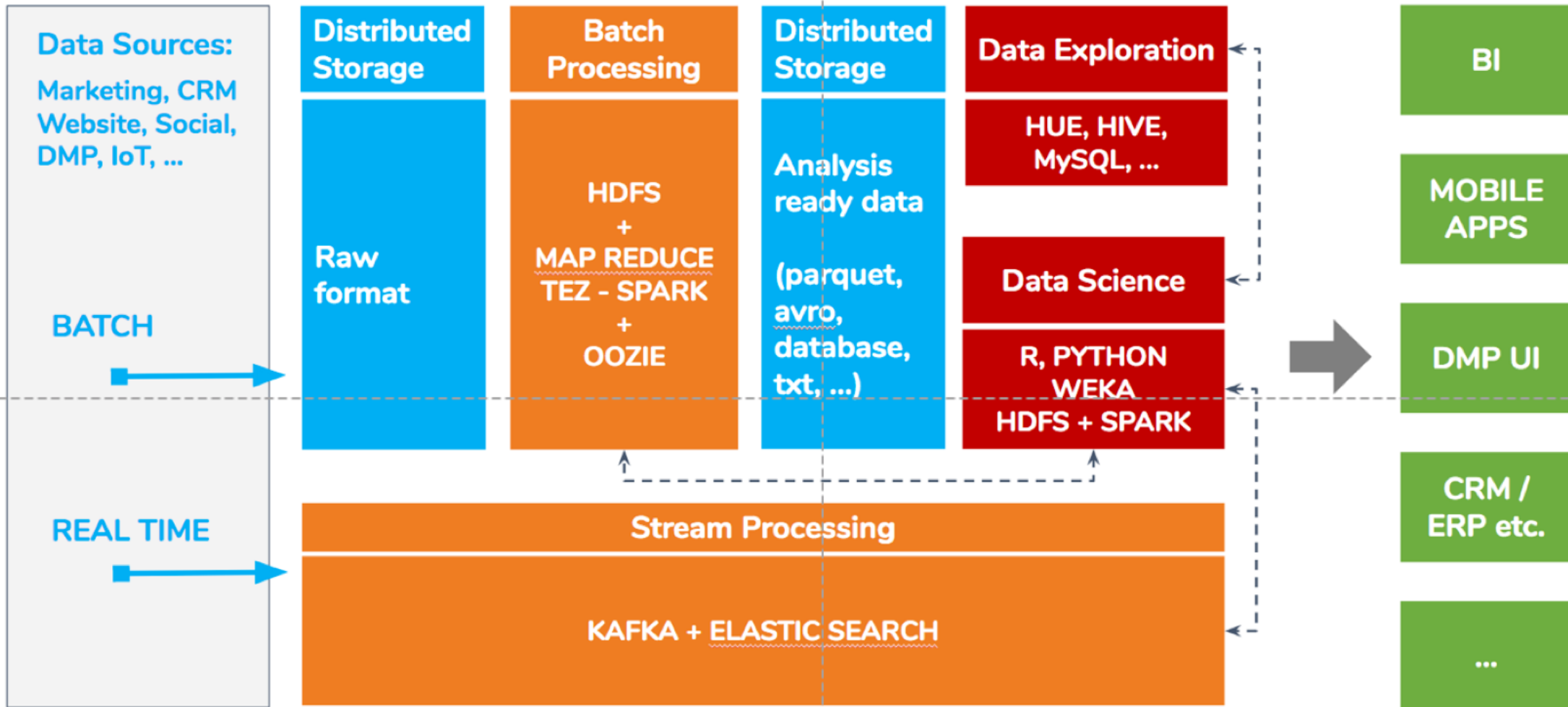
Data Extraction:

Per estrarre dati o liste utenti

Analisi prototipale:

Per sviluppare applicazioni specifiche

DATA LAKE – PLATFORM ARCHITECTURE



LEGENDA

- INGESTION & STORAGE
- PROCESSING
- ANALYSIS
- INTEGRATION