RED HAT
OPEN SOURCE DAY
Europe, Middle East & Africa

redhat.

L’architettura di Classe Enterprise di Nuova Generazione

Massimo Brignoli

Enterprise Account Executive

massimo@mongodb.com

@massimobrignoli

RED HAT
OPEN SOURCE DAY
Europe, Middle East & Africa

redhat.

Agenda

- Nascita dei Data Lake
- Overview di MongoDB
- Proposta di un’architettura EDM
- Case Study & Scenarios
- Data Lake Lessons Learned

mongoDB.
FOR GIANT IDEAS

RED HAT OPEN SOURCE DAY
Europe, Middle East & Africa

Quanti dati?

redhat.

Una cosa non manca alla aziende: dati

- Flussi dei sensori
- Sentiment sui social
- Log dei server
- App mobile

Analisti stimano una crescita del volume di dati del 40% annuo, 90% dei quali non strutturati.

Le tecnologie tradizionali (alcune disegnate 40 anni fa) non sono sufficienti

La Promessa del “Big Data”

Scoprire informazioni collezionando ed analizzando i dati porta la promessa di

- Un vantaggio competitivo
- Risparmio economico

Un esempio diffuso dell’utilizzo della tecnologia Big Data è la “Single View”: aggregare tutto quello che si conosce di un cliente per migliorarne l’ingaggio e i ricavi

Il tradizionale EDW scricchiola sotto il carico, sopraffatto dal volume e varietà dei dati (e dall’alto costo).

La Nascita dei Data Lake

Molte aziende hanno iniziato a guardare verso un'architettura detta Data Lake:

- Piattaforma per gestire i dati in modo flessibile
- Per aggregare I dati cross-silo in un unico posto
- Permette l'esplorazione di tutti i dati

La piattaforma più in voga in questo momento è Hadoop:

- Permette la scalabilità orizzontale su hardware commodity
- Permette una schema di dati variegati ottimizzato in lettura
- Include strati di lavorazione dei dati in SQL e linguaggi comuni
- Grandi referenze (Yahoo e Google in primis)

Perché Hadoop?

Hadoop Distributed FileSystem è disegnato per scalare su grandi operazioni batch

Fornisce un modello write-one read-many append-only

Ottimizzato per lunghe scansione di TB o PB di dati

Questa capacità di gestire dati multi-strutturati è usata:

- Segmentazione dei clienti per campagne di marketing e recommendation
- Analisi predittiva
- Modelli di Rischio

Ma va bene per tutto?

I Data Lake sono disegnati per fornire l'output di Hadoop alle applicazioni online. Queste applicazioni hanno dei requisiti tra cui:

- Latenza di risposta in ms
- Accesso random su un sottoinsieme di dati indicizzato
- Supporto di query espressive ed aggregazioni di dati
- Update di dati che cambiano valori frequentemente in real-time

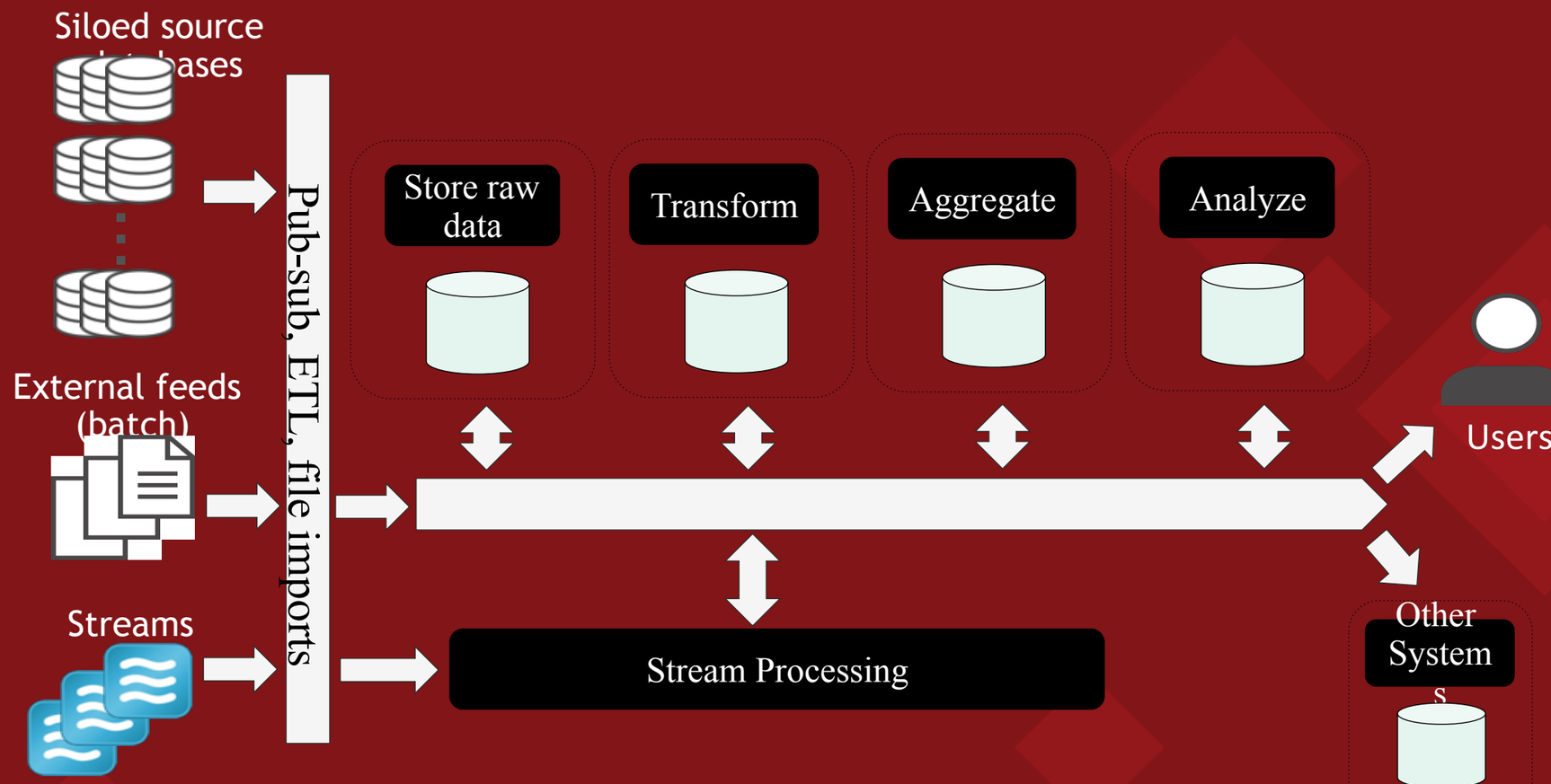
Hadoop è la risposta a tutto?

Nel nostro mondo guidato ormai dai dati, i millisecondi sono importanti.

- Ricercatori IBM affermano che il 60% dei dati perde valore alcuni millisecondi dopo la generazione
- Ad esempio identificare una transazione di borsa fraudolenta è inutile dopo alcuni minuti

Gartner predice che il 70% delle installazioni di Hadoop fallirà per non aver raggiunto gli obiettivi di costo e di incremento del fatturato.

Enterprise Data Management Pipeline



Stream icon from: https://en.wikipedia.org/wiki/File:Activity_Streams_icon.png

RED HAT
OPEN SOURCE DAY
Europe, Middle East & Africa

In Dettaglio

redhat.

- Join non necessarie causano pessime performance
- Costoso scalare verticalmente

- Lo schema rigido rende difficile il consolidamento di datai variabili o non strutturati
- Ci sono differenze nei record da eliminare durante la fase di aggregazione
- I processi soventi durano ore durante la notte

- I dati sono vecchi per prendere decisioni intraday

RED HAT
OPEN SOURCE DAY
Europe, Middle East & Africa

redhat.

Veloce Overview di MongoDB

RED HAT OPEN SOURCE DAY
Europe, Middle East & Africa

Documents Enable Dynamic Schema & Optimal Performance

redhat.

Relational

Customer ID	First Name	Last Name	City
0	John	Doe	New York
1	Mark	Smith	San Francisco
2	Jay	Black	Newark
3	Meagan	White	London
4	Edward	Daniels	Boston

Phone Number	Type	DNC	Customer ID
1-212-555-1212	home	T	0
1-212-555-1213	home	T	0
1-212-555-1214	cell	F	0
1-212-777-1212	home	T	1
1-212-777-1213	cell	(null)	1
1-212-888-1212	home	F	2

MongoDB

```
{  customer_id : 1,
   first_name : "Mark",
   last_name : "Smith",
   city : "San Francisco",
   phones: [
   {
   number : “1-212-777-1212”,
   dnc : true,
   type : “home”
   },
   number : “1-212-777-1213”,
   type : “cell”
   }]
}
```

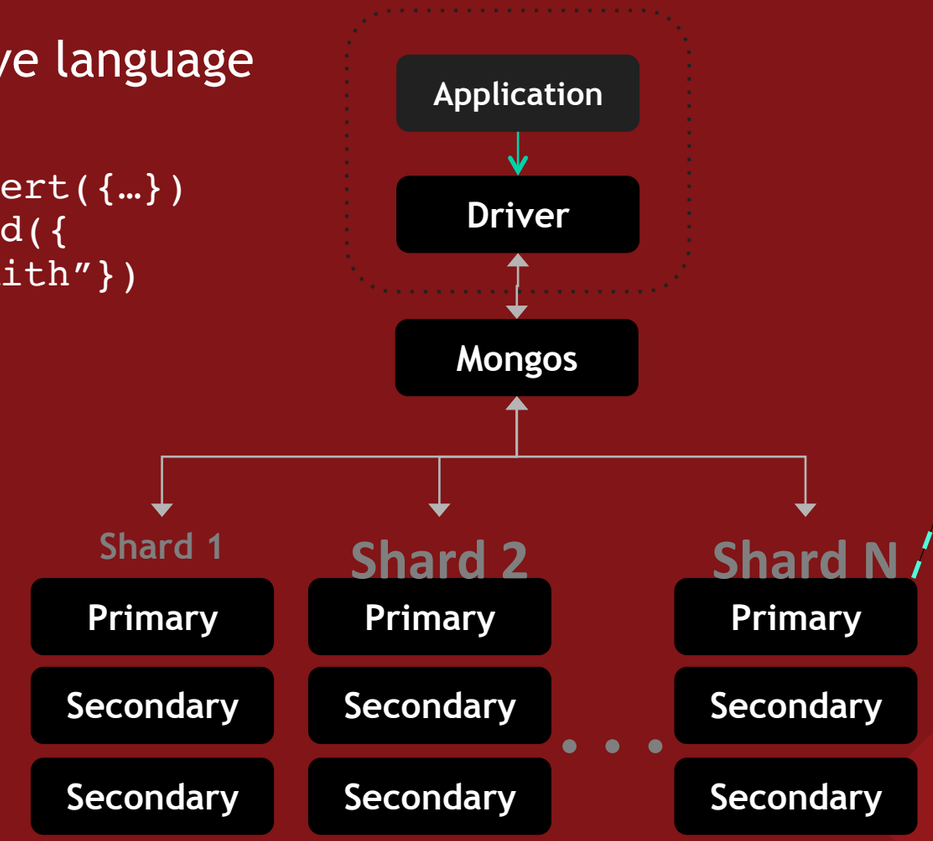
MongoDB Technical Capabilities

1. Dynamic Document Schema

```
{ name: “John Smith”,
  date: “2013-08-01”,
  address: “10 3rd St.”,
  phone: {
      home: 1234567890,
      mobile: 1234568138 }
}
```

2. Native language drivers

```
db.customer.insert({…})
db.customer.find({
 name: ”John Smith”})
```



3. High availability

4. Workload Isolation

5. High performance

- Data locality
- Indexes
- RAM

6. Horizontal scalability

- Sharding

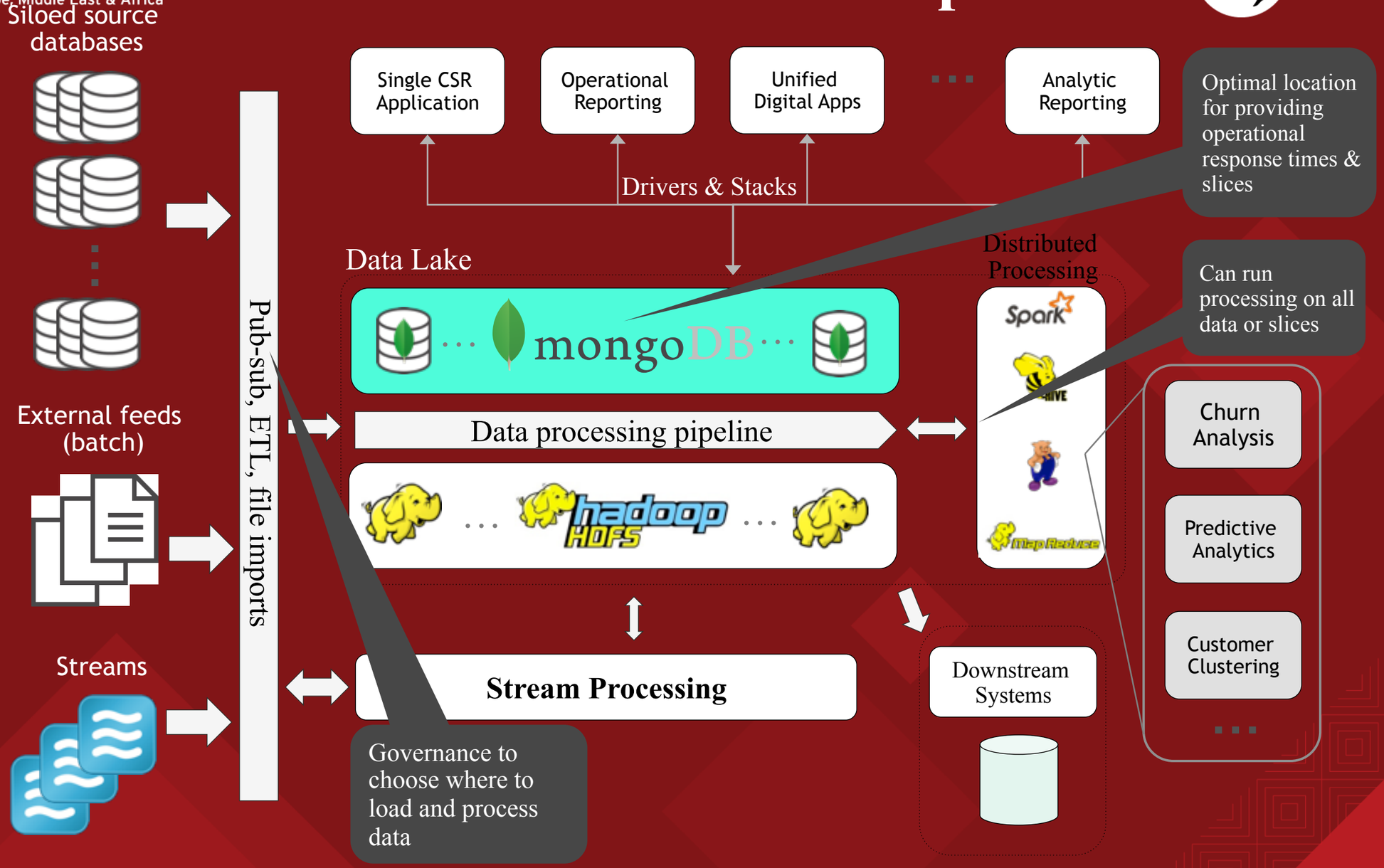
3.2 Features Relevant for EDM

- WiredTiger as default storage engine
- In-memory storage engine
- Encryption at rest
- Document Validation Rules
- Compass (data viewer & query builder)
- Connector for BI (Visualization)
- Connector for Hadoop
- Connector for Spark
- $lookUp (left outer join)

RED HAT OPEN SOURCE DAY
Europe, Middle East & Africa

Architettura EDM Completa

redhat.



Example scenarios

1. Single Customer View
    a. Operational
    b. Analytics on customer segments
    c. Analytics on all customers

2. Customer profiles & clustering

3. Presenting churn analytics on high value customers

RED HAT
OPEN SOURCE DAY
Europe, Middle East & Africa

redhat.

Grazie

Massimo Brignoli
Enterprise Account Executive
MongoDB Italia
#redhatosd