Integracja źródeł NoSQL

Arkadiusz Kasprzak
GFT
Krzysztof Pupiec
NASPERS
Nadia Sikorska
naviexpert
Marcin Siudziński
apollogic it KONTRAKT

Różnorodne źródła danych

Narzędzia Business Intelligence

Wybrane technologie

ORACLE

mongoDB

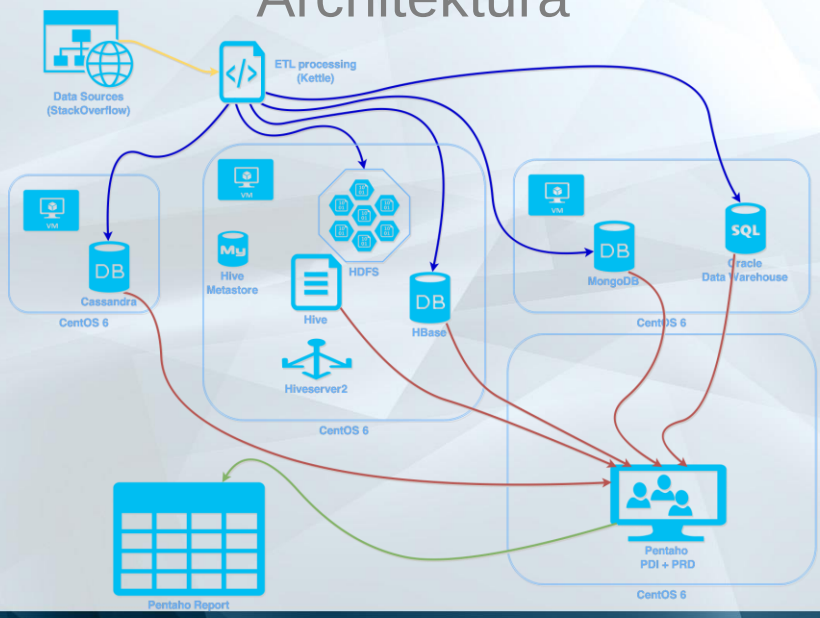
pentaho
A Hitachi Data Systems Company

HIVE

cassandra

APACHE HBASE

Architektura



Import danych

stackoverflow

Comment
- # Id
- ▫ PostId
- ◊ Score
- ◊ Text
- ◊ CreationDate
- ◊ UserId

PostLink
- # Id
- ◊ CreationDat
- ▫ PostId
- ◊ RelatedPostId
- ◊ LinkTypeId

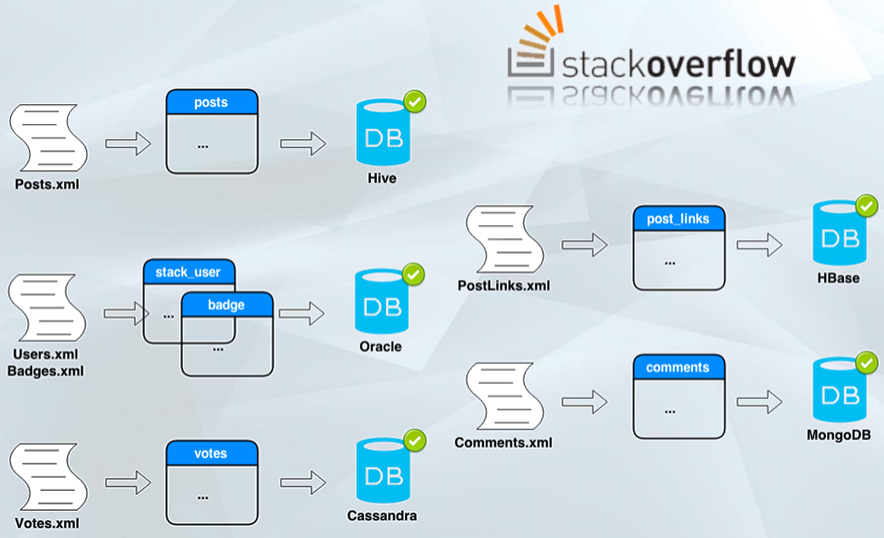
Vote
- # Id
- ▫ PostId
- ◊ VoteTypID
- ◊ CreationDate

Post
- # Id
- ◊ PostTypeId
- ◊ AcceptedAnswerId
- ◊ CreationDate
- ◊ Score
- ◊ ViewCount
- ◊ Body
- ▫ OwnerUserId
- ◊ LastEditorUserId
- ◊ LastEditorDisplayName
- ◊ LasteditDate
- ◊ LastActivityDate
- ◊ Title
- ◊ Tags
- ◊ AnswerCount
- ◊ CommentCount

User
- # Id
- ◊ Reputation
- ◊ CreationDate
- ◊ DisplayName
- ◊ LastAccessDate
- ◊ WebsiteUrl
- ◊ Location
- ◊ Views
- ◊ UpVotes
- ◊ DownVotes
- ◊ AccountId

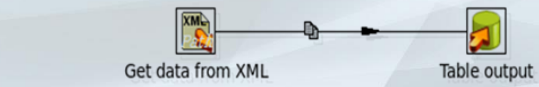
Badge
- # Id
- ▫ UserId
- ◊ Name
- ◊ Date

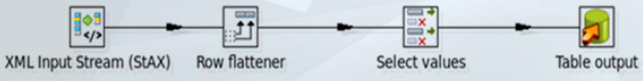
Import danych



Proces ETL

ORACLE

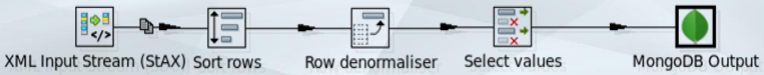




Proces ETL

mongoDB

✗brak wsparcia języka SQL

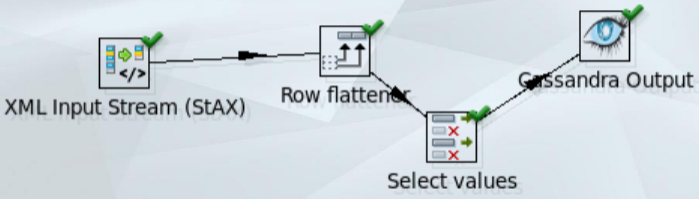


Proces ETL

cassandra

✗tabele muszą być kompatybilne wstecz - WITH COMPACT STORAGE

✗brak konektora PRD



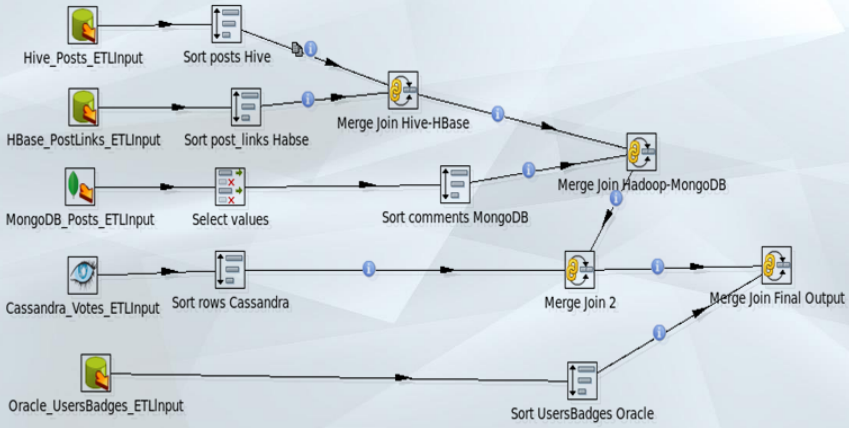
Proces ETL

- ✗brak konektora PRD
- ✗błędnie działający komponent w PDI
- ✗dostęp do bazy tylko przez Hive
- ✗wymagana dodatkowa konfiguracja sterownika „Generic Database”

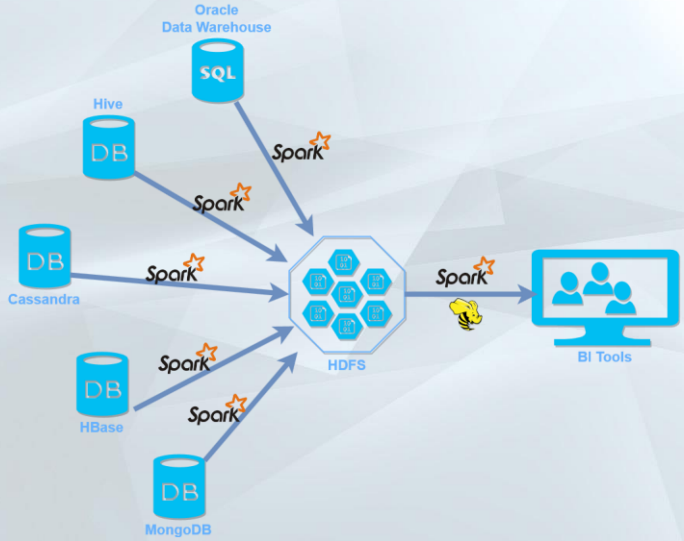
Proces ETL

- ✗niezgodność Pentaho Shims i Hadoop
- ✗problemy z obsługą distributed cache
- ✗brak wsparcia dla nowych wersji
- ✗błędnie działające komponenty w PDI
- ✗procesy ETL wymagają obsługi HDFS z poziomu OS

Konsolidacja źródeł



Rozwiązanie alternatywne



Cechy Apache Spark

- API w Java, Scala, Python, R
- Szybkość
- In-memory
- Równoległe operacje
- RDD (Resilient Distributed Datasets)
- Kompatybilność
- Odporność na awarie
- Dostępność bibliotek
- Strumieniowanie
- Algorytmy uczenia maszynowego

Apache Spark

Spark + Zeppelin

Dziękujemy za uwagę.

Arkadiusz Kasprzak
Krzysztof Pupiec
Nadia Sikorska
Marcin Siudziński