New Database Replication and Data Integration with Hadoop and BI

Jeffrey Surretsky
NYOUG
December 2013
Big Data – Hadoop®
The explosion of data continues to burden the data tool chain

Transactional Data
Traditionally, only transactional data was generated and stored in databases
- Structured
- Measured growth

Human Files
But over time, we started creating unstructured data
- Docs, Images, Video
- Multiple formats
- Fast growth

Social & Machines
have added exponentially
- Likes, tweets, relationships (social)
- Log files (machine)
- Exponential growth

mainframe ---- PC ---- internet ---- mobile ---- machine ---->
Big data market drivers

- Proliferation of new user generated data creation and data capture technologies
- Increased “interconnectedness” drives consumption (creating more data)
- Inexpensive storage makes it possible to keep more data longer
- Need to extract actionable insights from all data assets to gain competitive edge

*Source: IDC 2011*
Big data

Scaling up on RDBMs

• Partitioning
• Materialized Views
• In memory cache
• ...and who are we kidding here!

RDBMS Yodabytes handle cannot!
Big data
RDBMS Cluster
Big data - Hadoop
Big data – Hadoop benefits

- Scalable storage
- Massive parallel processing
- Cost effective
Hadoop operational use cases

1. Staging
2. Archiving
3. Warehousing

Not glamorous, but highly effective.
Today’s solutions

OLTP

Data Warehouse

Analytics
Log-based CDC Replication

- Near real-time log-based CDC from Oracle
- Applying Changes to Hadoop
Log-based CDC from Oracle-to-Oracle Architecture

Source

Redo/Archive logs

Capture queue

Capture

Read

Export queue

Export

Import

Post queue

SQL

Post

Target
Log-based CDC Replication – impact-free and limitless!

- Operational reporting/Archiving/Data warehousing
- Change tracking
- Cascading using intermediary systems
- Data distribution/Distributed processing
- Data integration
- Migrations, patches & upgrades
- Centralized reporting (Consolidation)
- Load balancing
- High availability/Disaster recovery
Log-based CDC Data Integration Architecture

Combined source & target process implementation

Redo/Archive logs

Oracle source

Capture queue

Post queue

Capture

Read

Near real-time data integration

JMS post

JMS queue

Custom App

Target(s)

Dell App

Oracle source

Repose/Archive logs

Capture queue

Post queue

Capture

Read
Log-based CDC Database Replication & Near Real-time Data Integration Summary

Database replication

Near real-time data integration

Source

Target(s)

JMS queue

Custom app

…And more
Connector for Hadoop

• Provides **near real-time** data replication from Oracle to Hadoop environments. The solution enables organizations to affordably replicate live data from Oracle tables
  – In near real time to HDFS and Hive environments
  – In real time to HBase
SQOOP

ORACLE

HBase

HDFS

hadoop
Log-based CDC \rightarrow JMS \rightarrow Connector for Hadoop

HBase \rightarrow HDFS

ORACLE

Hadoop

HDFS

JMS
SharePlex for Oracle → JMS → Connector for Hadoop

HBase → HDFS
SharePlex for Oracle → JMS → Connector for Hadoop

ORACLE

HBase

HDFS

hadoop
Log-based CDC → JMS → Connector for Hadoop

HBase ➔
HDFS ➔
Log-based CDC → JMS → Connector for Hadoop

HBase

HDFS
Log-based CDC → JMS → Connector for Hadoop

HBase

HDFS
Log-based CDC → JMS → Connector for Hadoop

HBase

HDFS
Log-based CDC → JMS → Connector for Hadoop → HBase → HDFS
Log-based CDC → JMS → Connector for Hadoop

HBase → HDFS
Log-based CDC → JMS → Connector for Hadoop
Log-based CDC → JMS → Connector for Hadoop → HBase → HDFS
SharePlex Connector for Hadoop architecture

Log-based CDC → SQOOP → HBase

JMS → HDFS

Connector for Hadoop
SharePlex Connector for Hadoop – use case

- Oracle Financials
- SAP Manufacturing
- PeopleSoft HR
- Siebel CRM

Data warehouse, stage and archive

Reporting

Dashboards

Analytics
Questions
The power to do more