



Bridging the Big Data Divide with Oracle Data Integration

*Milomir Vojvodic,
Business Development Manager, EMEA DIS*



98,000+
TWEETS



370,000+ MINUTES
VOICE CALLS ON
skype



320+
NEW
twitter
ACCOUNTS

100+
NEW
Linked in
ACCOUNTS

1 associated content
NEW
ARTICLE IS
PUBLISHED

Y!
THE
WORLD'S
LARGEST
COMMUNITY
CREATED CONTENT!

6,600+
NEW
PICTURES ARE
UPLOADED ON
flickr



50+
WORDPRESS
DOWNLOADS



125+
PLUGIN
DOWNLOADS

79,364
WALL
POSTS

510,040
COMMENTS



695,000+
facebook
STATUS
UPDATES

1,700+
Firefox
DOWNLOADS



Google

Google Search

694,445
SEARCH
QUERIES

168 MILLION
EMAILS
ARE SENT



60+
NEW
BLOGS

1,500+
BLOG
POSTS

70+
DOMAINS
REGISTERED



600+
NEW
VIDEOS

100+
Answers.com
40+
YAHOO! ANSWERS

QUESTIONS
ASKED ON THE
INTERNET...

25+ HOURS
TOTAL
DURATION



20,000+
NEW
POSTS ON
tumblr.



13,000+
iPhone
APPLICATIONS
DOWNLOADED



1,600+
READS ON
Scribd.



THE
LARGEST
SOCIAL READING
PUBLISHING COMPANY

13,000+ HOURS
MUSIC
STREAMING ON
PANDORA



12,000+
NEW ADS
POSTED ON
craigslist



1 **NEW**
DEFINITION
IS ADDED ON
urban



Diverse Data Sets

Information Architectures Today:
Decisions based on transactional data



transactions, applications, structured Data

Information Architectures Today:
Decisions based on all your data

Video and Images



Documents

Social Data



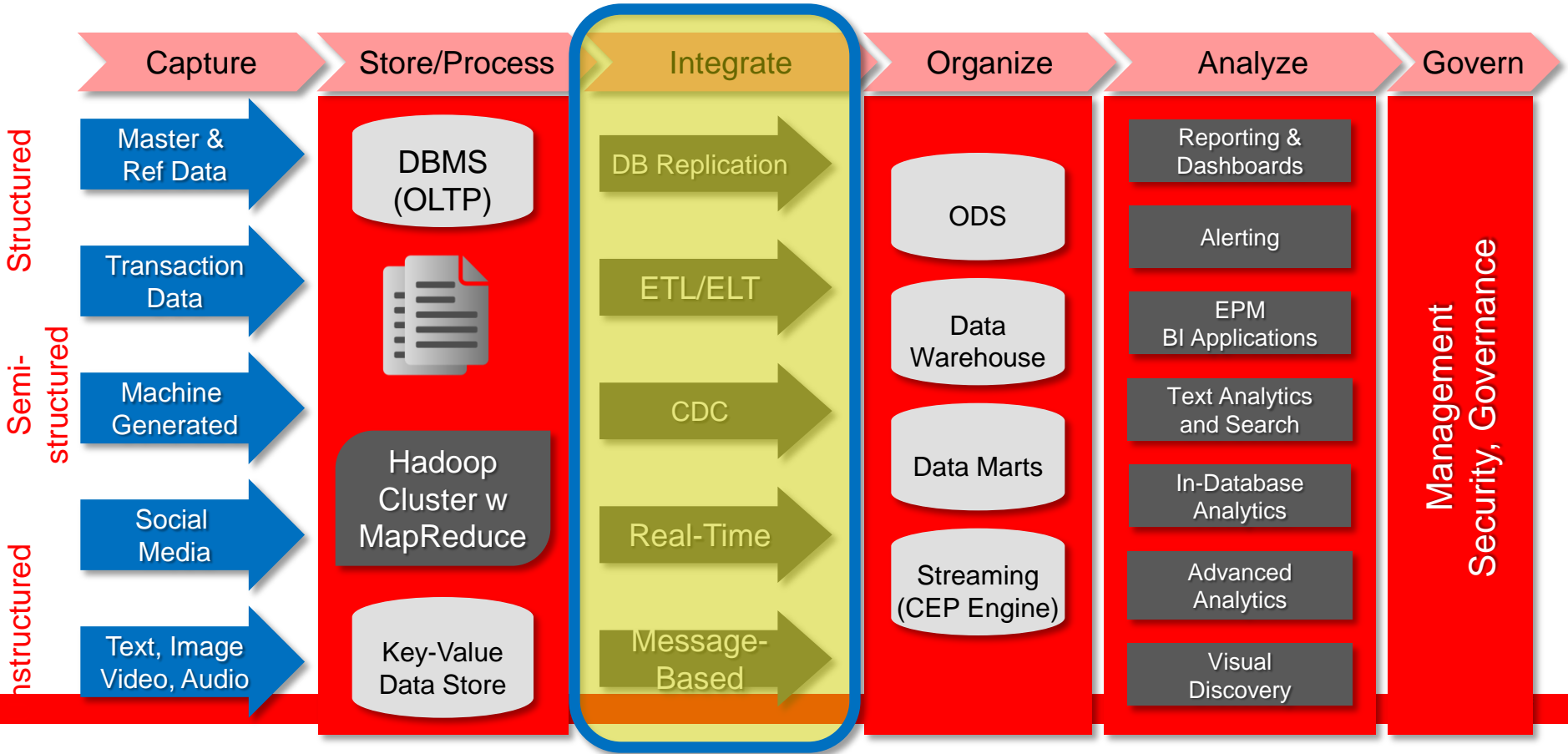
Machine-Generated Data



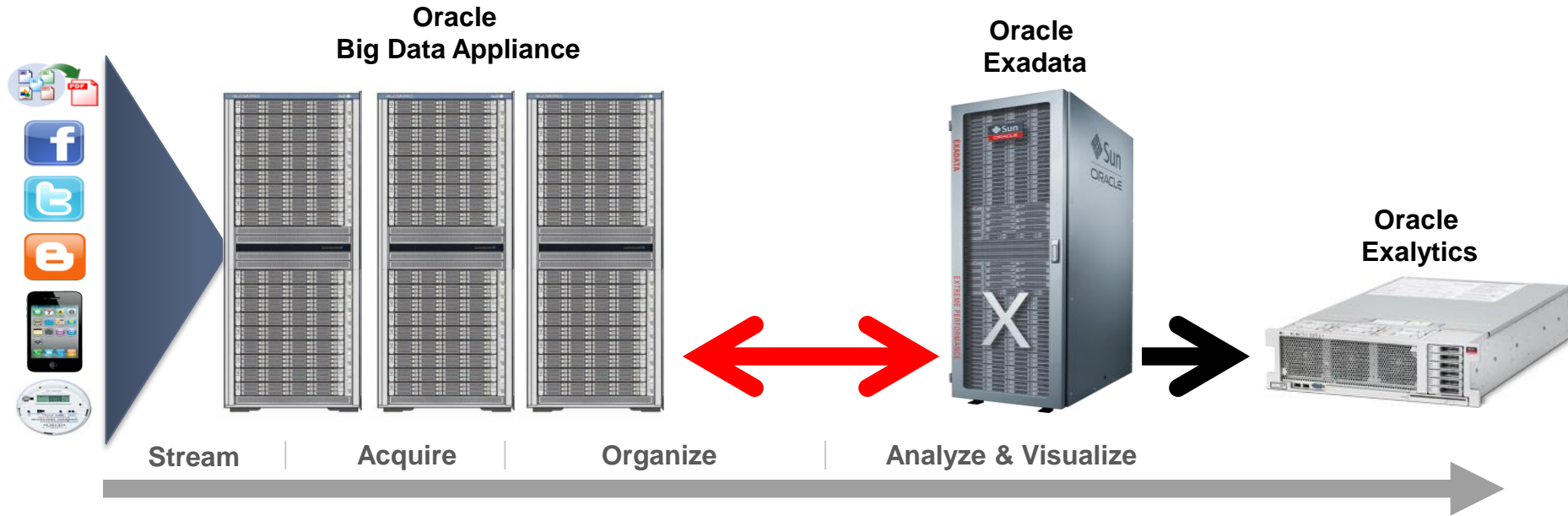
Oracle Data Integration Solutions for Big Data



Integrated Architecture

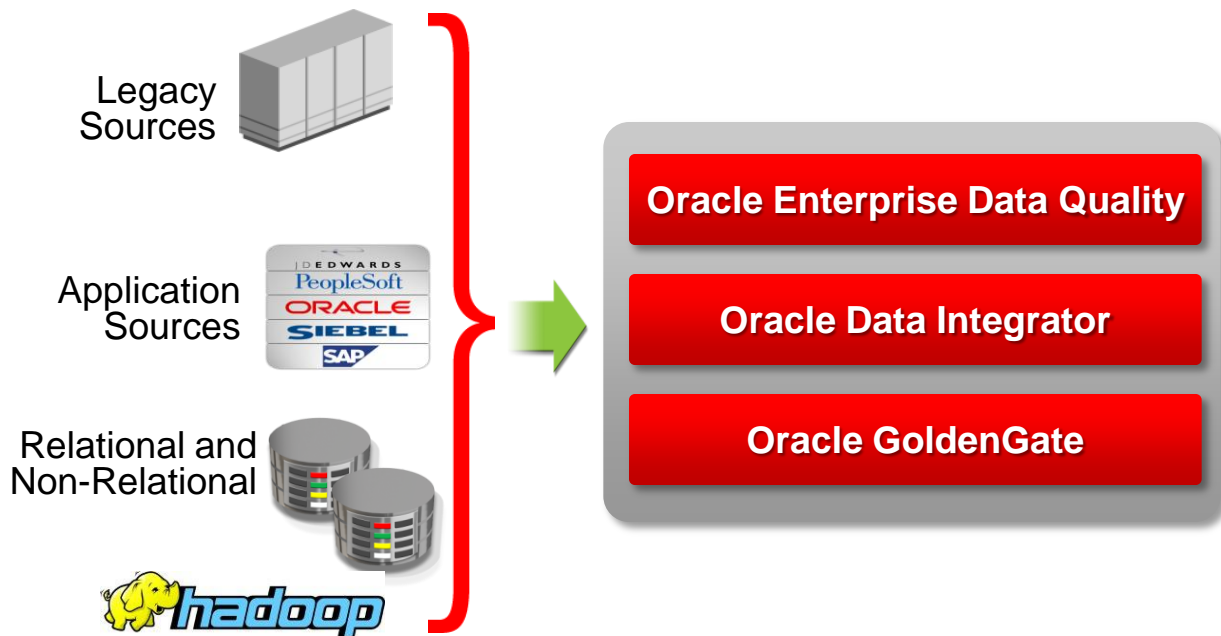


Integrate Big Data with DW and Transactional Data Stores



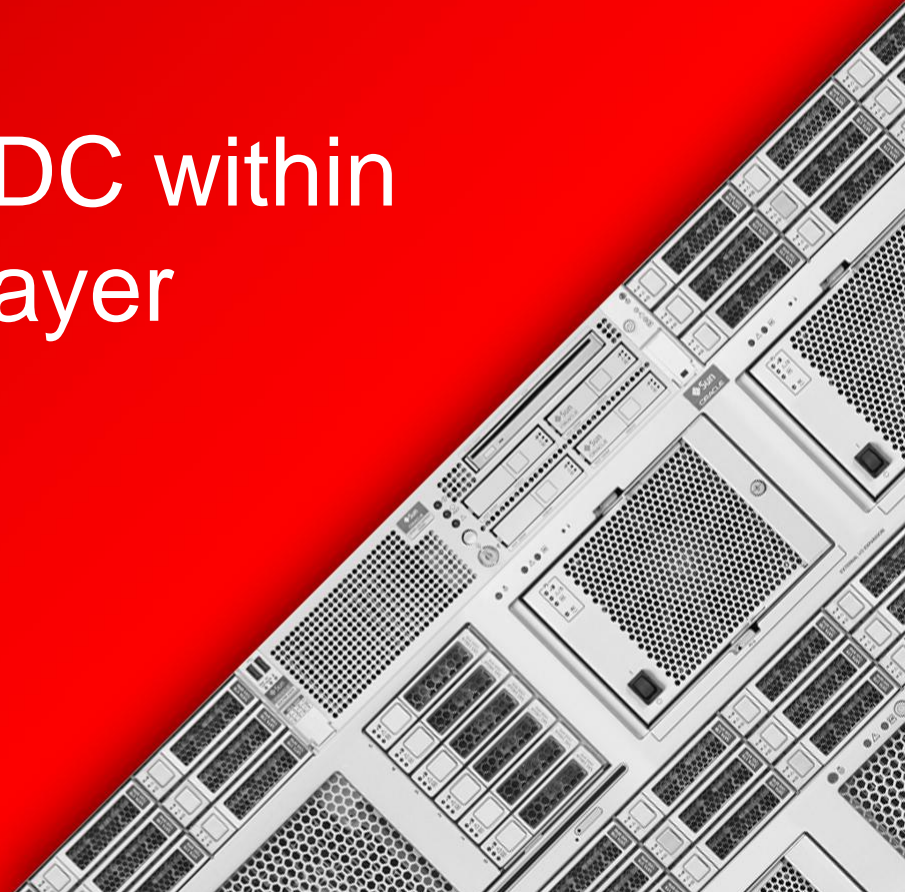
- ❑ Load from big data processing into your data warehouse for further analysis
- ❑ Access your customer information while you process through your big data in order to look for patterns

Oracle Data Integration Solutions



- Complete and best-of-breed approach to address **enterprise integration**
- **Maximum performance** with lower cost of ownership, ease of use, and reliability.
- Certified for leading technologies to deliver **fast time to value**
- Oracle customers report:
 - 80% lower TCO
 - Five times higher performance
 - 70% reduction in development costs

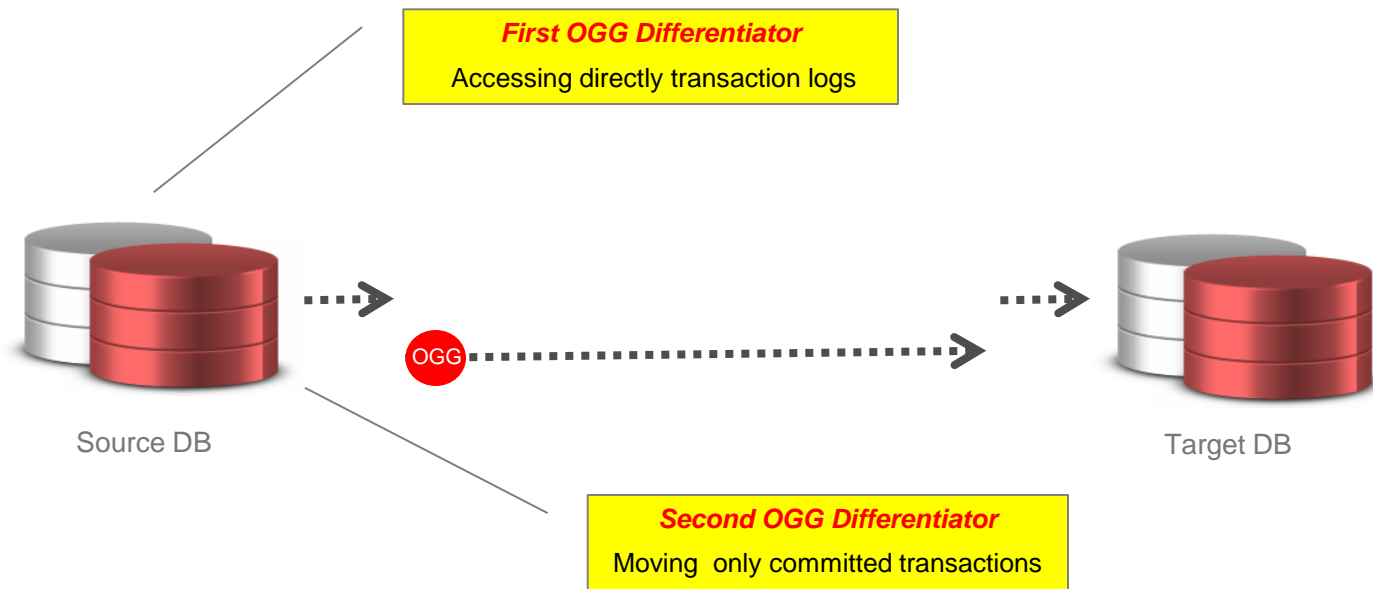
DB Replica and CDC within Data Integration Layer



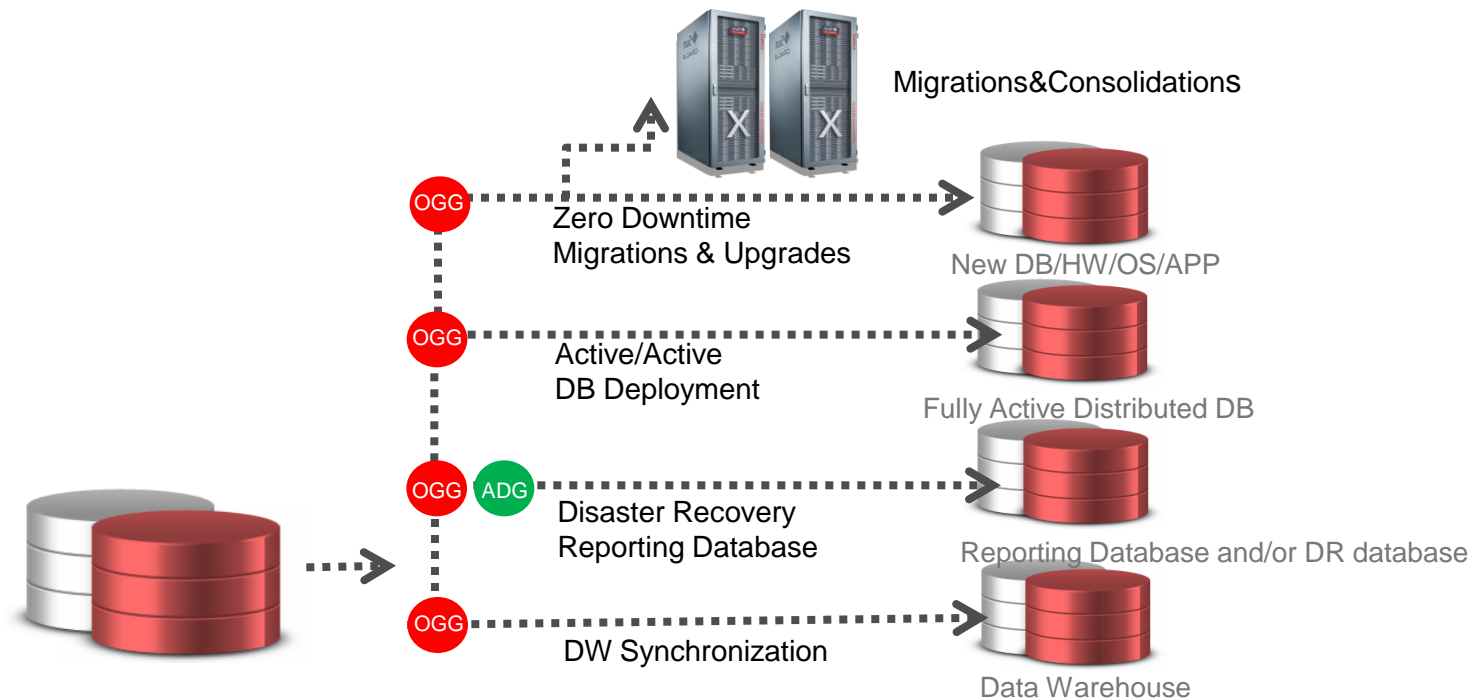
What is Oracle GoldenGate?



What is Oracle GoldenGate?



Oracle DIS Use Cases - OGG



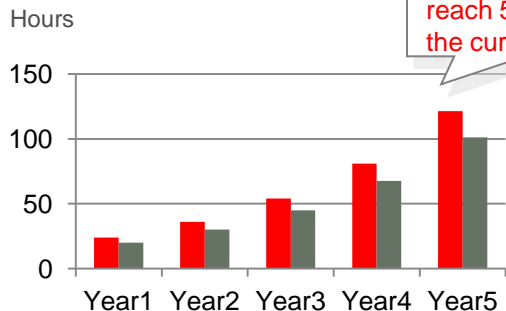
OGG is Log Based Replica

OR

Comparing to round 50% of cases :

Scanning of the whole DB
(batch window)

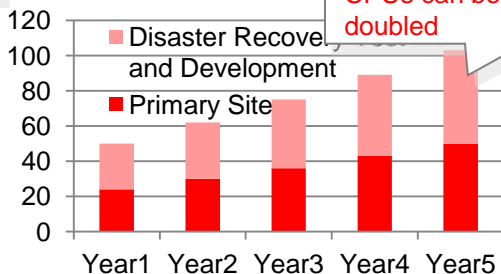
TIME REQUIRED FOR THE END OF DAY PROCEDURE



Currently during the End Of Day utilizes the Server CPU by 40-50% and the IO by 90%. Probably the IO is the bottleneck.

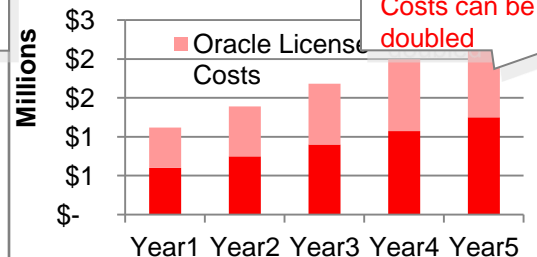
NO OF CPUs REQUIRED FOR PERFORMANCE*

No Of Required CPUs

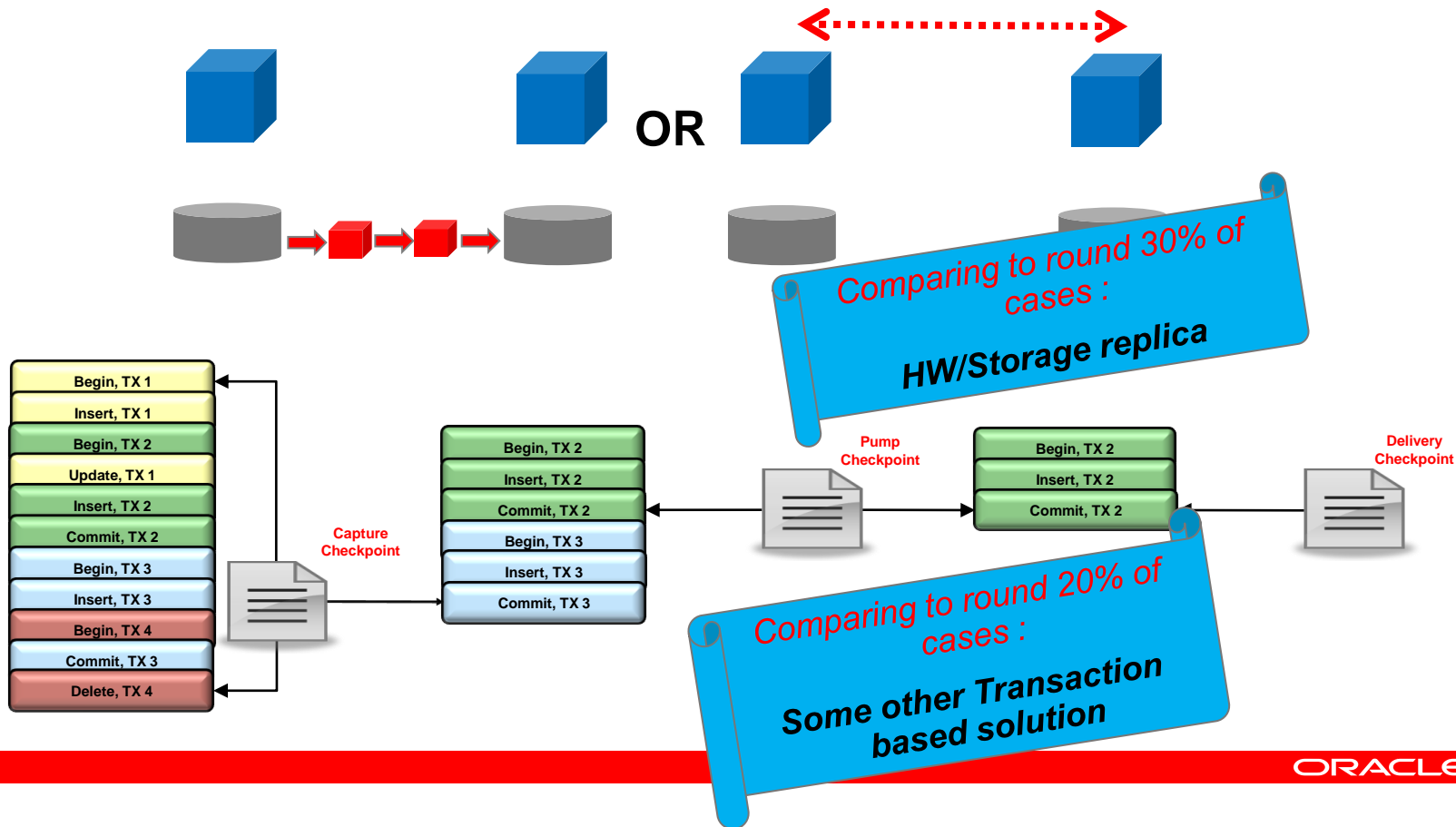


ESTIMATED COSTS FOR SERVER AND LICENSE**

Estimated Cost of Purchase in USD



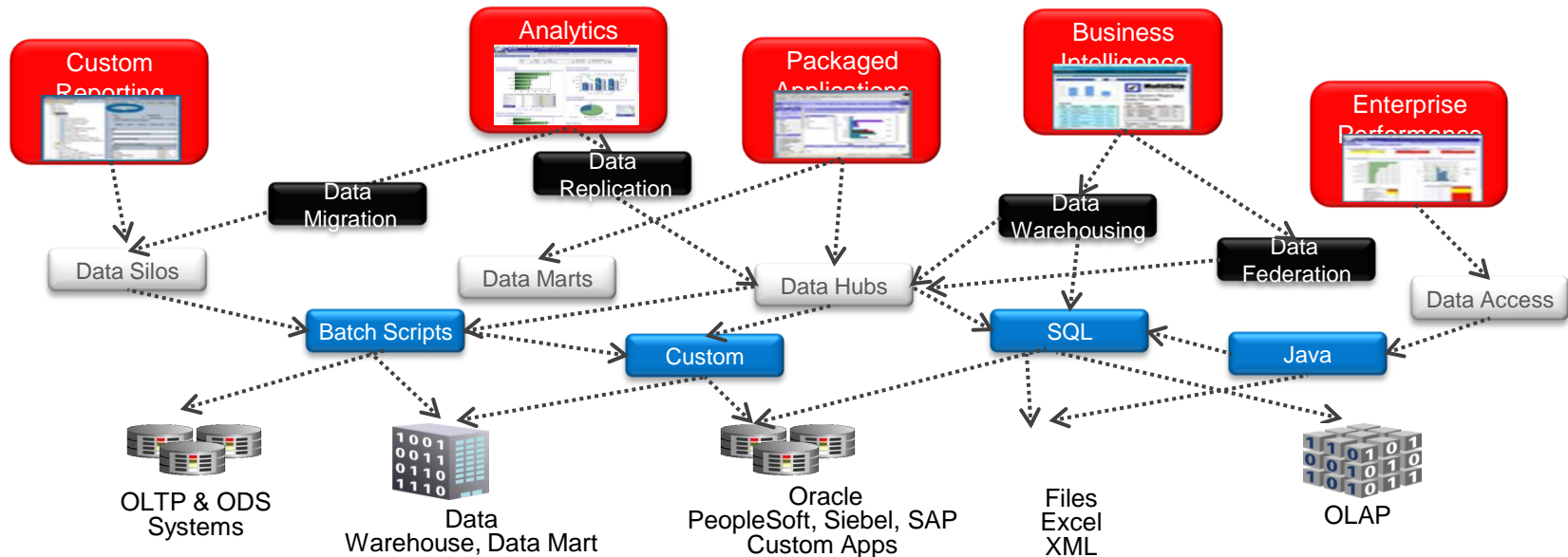
OGG Moves Only Committed Transactions



ETL and Data Quality within Data Integration Layer



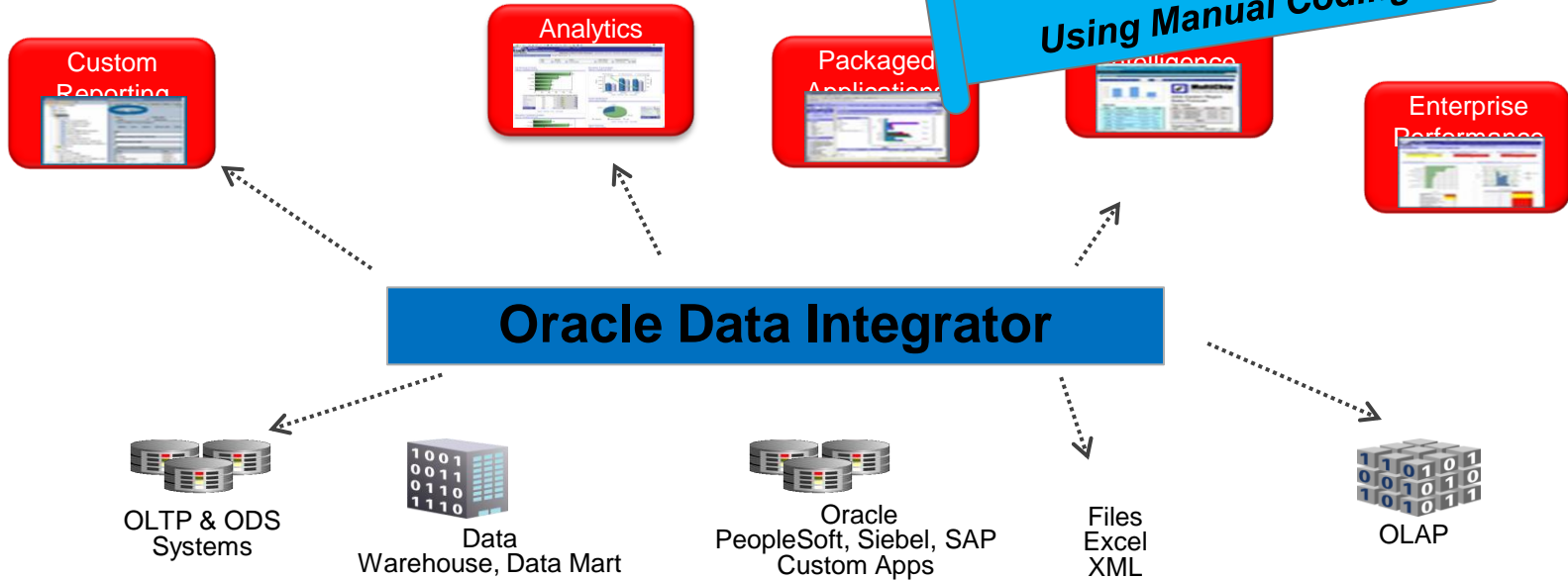
ODI is centralizing all ETL Development



ODI is centralizing all ETL Development

Comparing to round 80% of cases :

Using Manual Coding

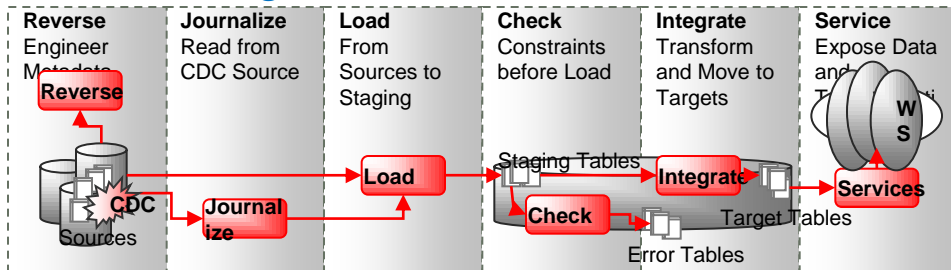


Why is ODI different?

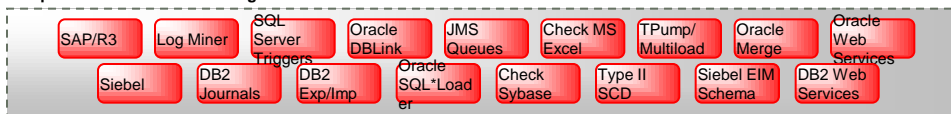
Second ODI Differentiator

ODI Declarative Design and ODI Knowledge Modules for reusing already written down level SQL code

ODI Knowledge Modules

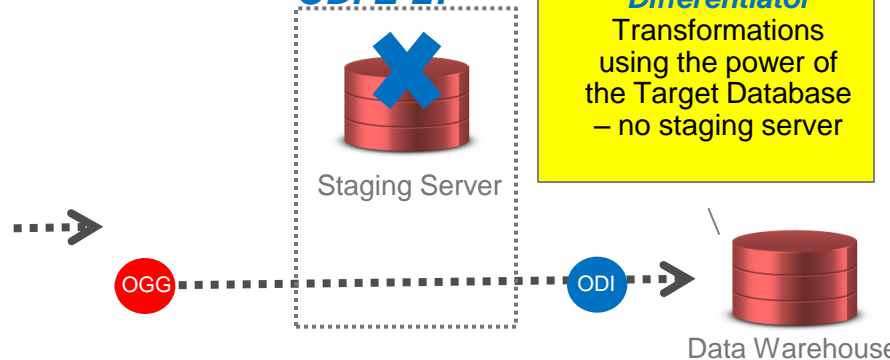


Sample out-of-the-box Knowledge Modules



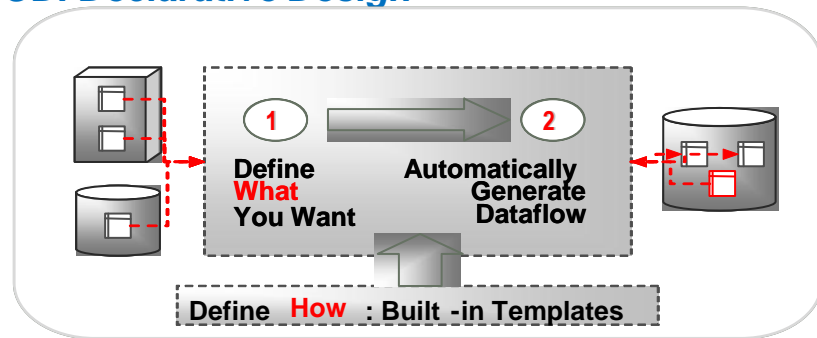
Benefits

ODI E-LT



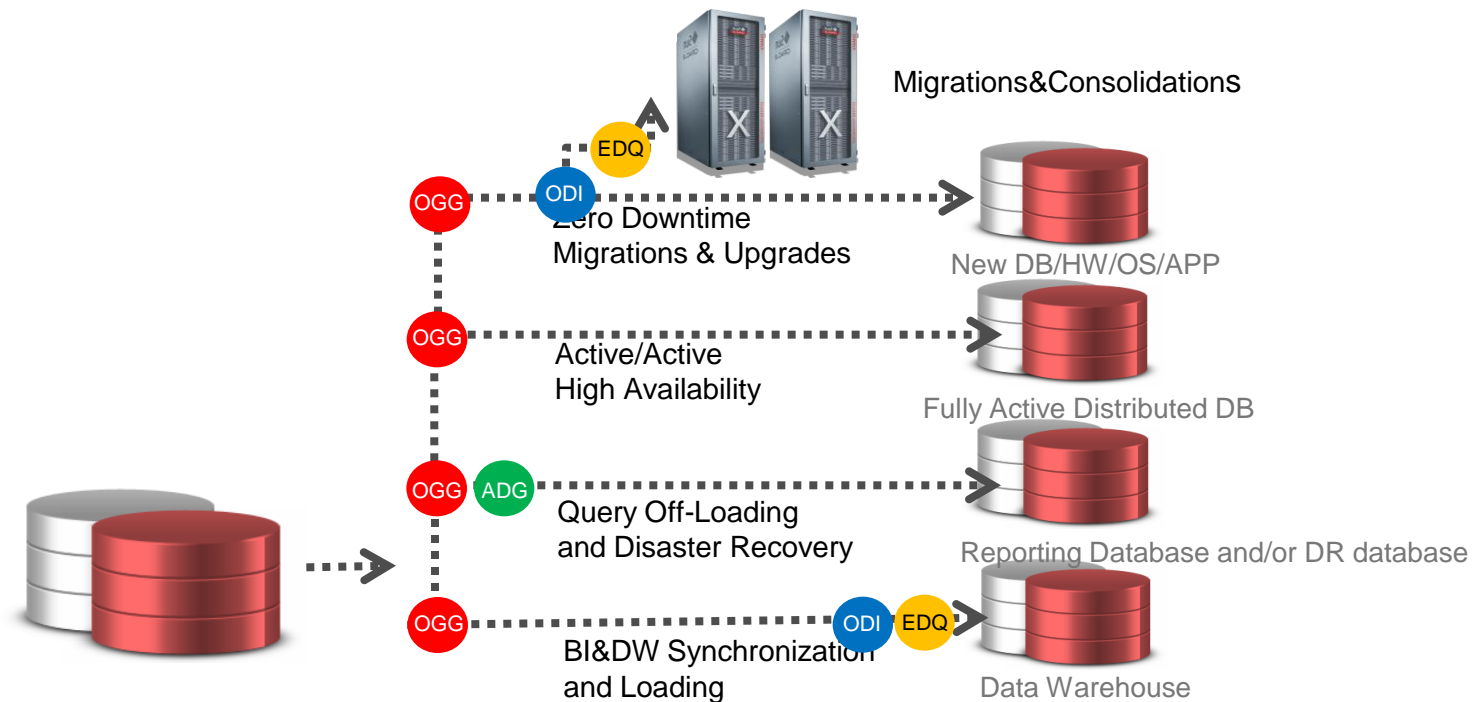
First ODI Differentiator
Transformations using the power of the Target Database – no staging server

ODI Declarative Design



ORACLE

Oracle DIS Use Cases – ODI and EDQ



Why Do We Need Data Quality?

Inconsistent formats

Abbreviations
(often ambiguous)

Attributes non-standard,
missing or invalid

Customer ID	Customer Name	Address 1	Address 2	City	State	Zip	Country	Birth Date	Gender
AD23298	Mr Peter Mayhew	9407 Main St		Fairfax	VA	22031-4001	USA	02/23/61	M
VS38611	Dr Ellen Van Der Heijde 1	144 E Grove St		Kingston	PA	18704	US	07/12/57	
DC18223	Jalila Abdul-Alim (Do Not Call) 2	4548 Pennsylvania Ave	Apt 205	Kansas City	MO	64111-3349	USA	02/23/63	F
CO9387A	Tayside Computers Inc. 3	4912 E 41st N		Idaho Falls	ID	83401	USA	31/03/2007 7	
TZ35019	Mr Zachary P Jahn	98-1731 Ipuala Loop	Aiea	Hawaii	96701 5		United States	06/12/86	Male
CB27843	Mrs Edith Y Baba Junior	Baba Real Est. Corp. 3	209 Stony Point Trl	Webster	NY		USA	11/17/1971	M 6
OX80306	Andrew & Mary Baxter 4	14 Oxbridge Way		Milford 6	NH	03055-4614	US	05/28/67	F
JP70210	Mr RJ & Mrs FB MacDonald 4	57 Hadleigh Close	Westlea	Swindon SN5 9BZ	MA				Y
RD48107	Mr Andy Baxter	14 Oxbridge Wy		Milford	NH	3056	USA	01/01/01 8	M

Widespread
duplication
(often hard
to spot)

1

Compound Names

2

Embedded Additional Information

3

Mixed Business & Personal Names

4

Multiple Names

5

Mis-Fielded Data

6

Erroneous Data

7

International Date Formats

8

Default or Dummy Data

Why Do We Need Data Quality?



10hp motor 115V Yoke mount

MOT-10,115V, 48YZ,YOKE

mtr, ac(115) 10 horsepower 115volts

This 10hp yoke mounted motor is rated for 115V with a 5 year warranty

10 Caballos, Motor, 115 Voltios

TEAO HP = 10.0 1725RPM 115V 48YZ YOKE MTR

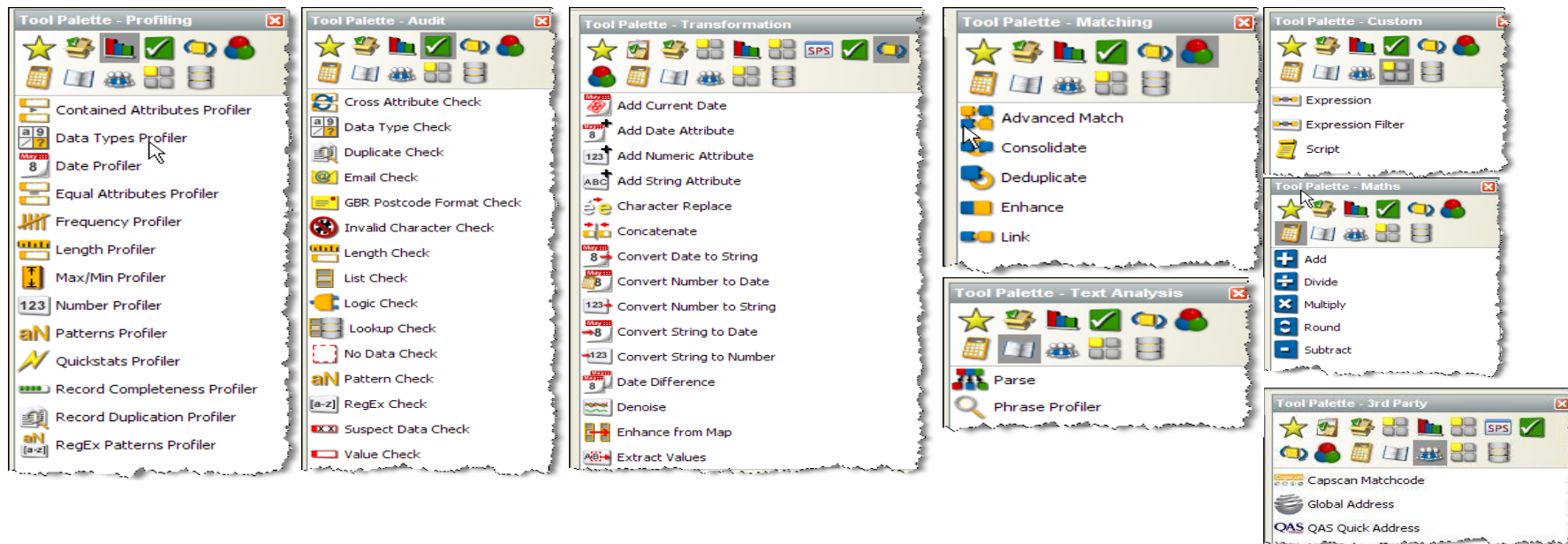
Motor, TEAO, 1725 RPM, 48YZ, 15 Voltios, Montaje de Yugo, hp = 10

Item	Motor
Classification	26101600
Power	10 horsepower
Voltage	115
Mounting	Yoke

Product data is much more variable and unpredictable than other data types

Oracle Enterprise Data Quality

- Profile, Audit, Transform, Parse, Cleanse, Standardize, Match within One Unified Solution*



EDQ Address Verification

300 Berry #1210 SF California

Latitude 37.775837
Longitude -122.39557

	Parse	Validate
PremiseNumber	300	300
ThoroughfareName	Berry	Berry St
SubPremise	#1210	Unit 1210
Locality	SF	San Francisco
AdministrativeArea	California	CA
PostCode		94158-1670

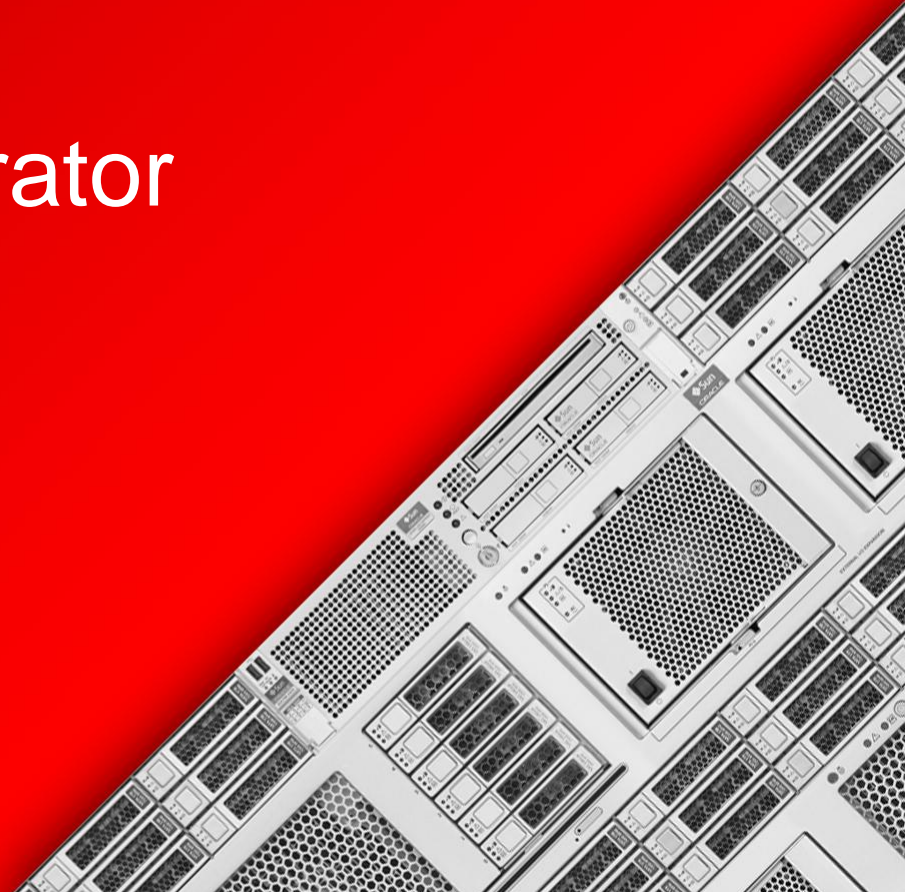
Step 1 Extract pieces of the address

Step 2 Check the pieces against the information in the Global Knowledge Repository to complete and find the correct abbreviations

Step 3 Change character set – transliterate - if necessary

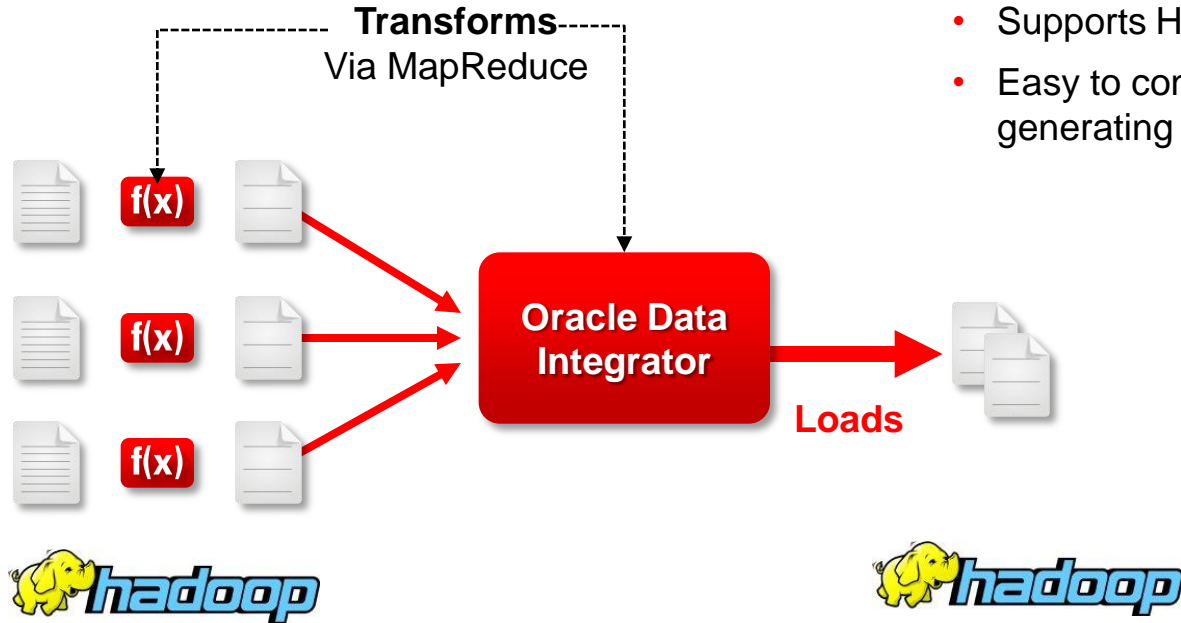
Step 4 Find Location

Oracle Data Integrator for Big Data



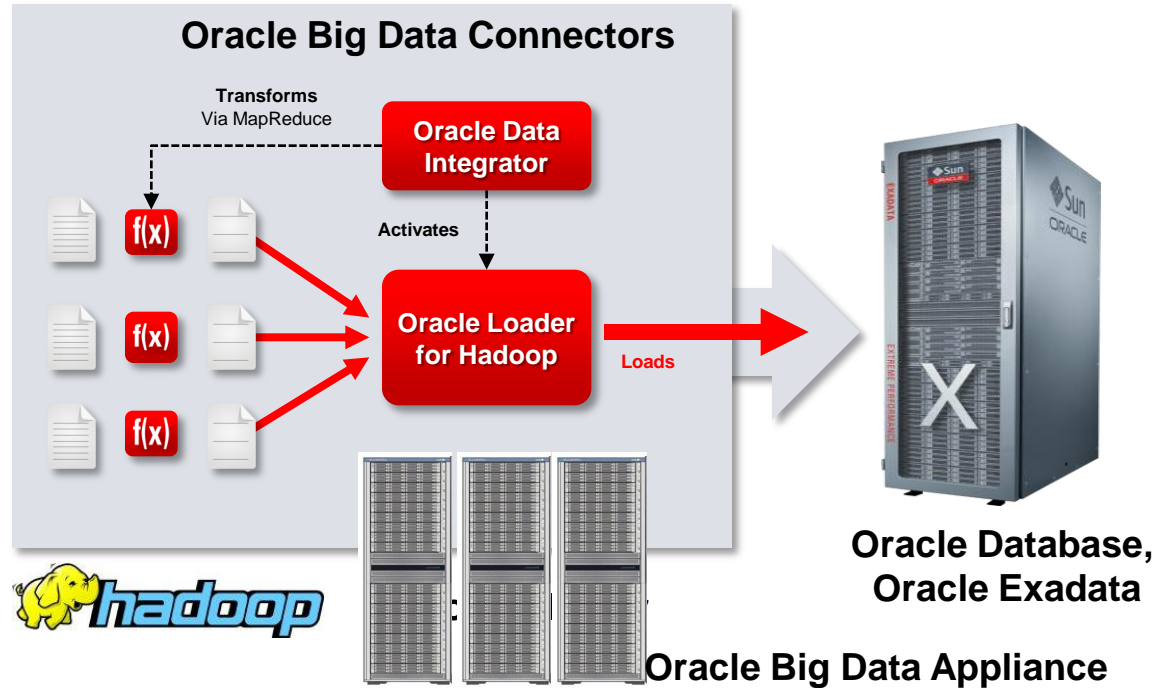
ODI for Big Data

Heterogeneous Integration to Hadoop Environments



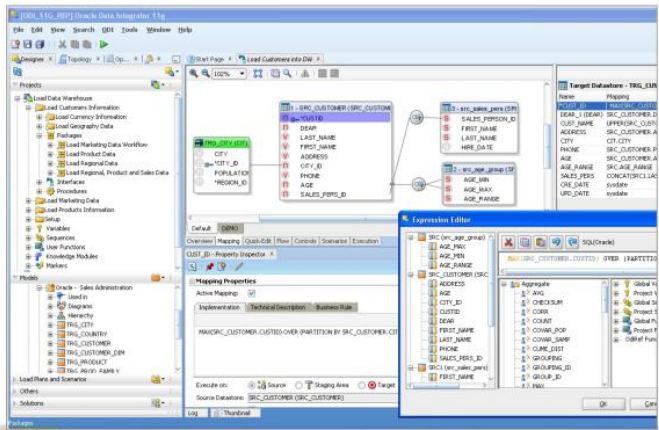
ODI for Big Data to Oracle

Optimized Integration to Oracle Exadata



Oracle Data Integrator for Big Data

Putting Together the Unique Advantages



Simplifies creation of Hadoop and MapReduce code to boost productivity

Integrates big data heterogeneously via industry standards: Hadoop, MapReduce, Hive, NoSQL, HDFS

Unifies integration tooling across unstructured/semi-structured and structured data

Optimizes loading of big data to Oracle Exadata using Oracle Big Data Connectors

Engineered for running on and integrating with Oracle Big Data Appliance via Big Data Connectors

