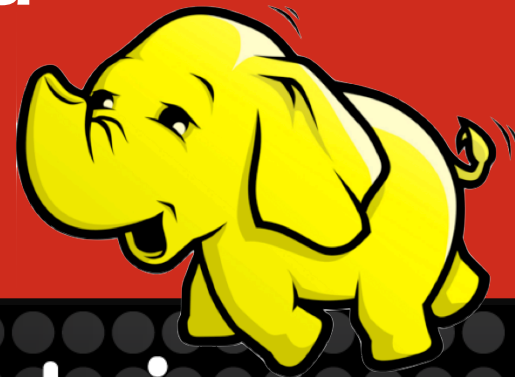


ORACLE®

Platinum
Partner

Building the Integrated Data Warehouse

with Oracle Database and Hadoop



Gwen Shapira, Senior Consultant

Pythian
love your data

13 years with a pager
Oracle ACE Director
Oak table member
Senior consultant for Pythian
@gwenshap
[http://www.pythian.com/
news/author/shapira/](http://www.pythian.com/news/author/shapira/)
shapira@pythian.com



Pythian

Recognized Leader:

- Global industry-leader in remote database administration services and consulting for Oracle, Oracle Applications, MySQL and Microsoft SQL Server
- Work with over 165 multinational companies such as LinkShare Corporation, IGN Entertainment, CrowdTwist, TinyCo and Western Union to help manage their complex IT deployments

Expertise:

- One of the world's largest concentrations of dedicated, full-time DBA expertise. Employ 7 Oracle ACEs/ACE Directors. Heavily involved in the MySQL community, driving the MySQL Professionals Group and sit on the IOUG Advisory Board for MySQL.
- Hold 7 Specializations under Oracle Platinum Partner program, including Oracle Exadata, Oracle GoldenGate & Oracle RAC

Agenda

- What is Big Data?
- Why do we care about Big Data?
- Why your DWH needs Hadoop?
- Examples of Hadoop in the DWH
- How to integrate Hadoop into your DWH
- Avoiding major pitfalls



ORACLE

Platinum
Partner

What is Big Data?

Pythian
love your data

Doesn't Matter.

We are here to discuss architectures.
Not define market segments.

What Does Matter?

Some data types are a bad fit for RDBMS.
Some problems are a bad fit for RDBMS.

We can call them BIG if you want.
Data Warehouses have always been BIG.

Given enough skill and money –
Oracle can do anything.

Lets talk about efficient solutions.

When RDBMS Makes no Sense?

- Storing images and video
- Processing images and video
- Storing and processing other large files
 - PDFs, Excel files
- Processing large blocks of natural language text
 - Blog posts, job ads, product descriptions
- Processing semi-structured data
 - CSV, JSON, XML, log files
 - Sensor data

When RDBMS Makes no Sense?

- Ad-hoc, exploratory analytics
- Integrating data from external sources
- Data cleanup tasks
- Very advanced analytics (machine learning)

New Data Sources

- Blog posts
- Social media
- Images
- Videos
- Logs from web applications
- Sensors

They all have large potential value

But they are awkward fit for traditional data warehouses



Your DWH needs Hadoop

Pythian
love your data

Big Problems with Big Data

- It is:
 - Unstructured
 - Unprocessed
 - Un-aggregated
 - Un-filtered
 - Repetitive
 - Low quality
 - And generally messy.

Oh, and there is a lot of it.

Technical Challenges

- Storage capacity
 - Storage throughput
 - Pipeline throughput
 - Processing power
 - Parallel processing
 - System Integration
 - Data Analysis
- } Scalable storage
- } Massive Parallel Processing
- } Ready to use tools

Hadoop Principles

Bring Code to Data



Share Nothing

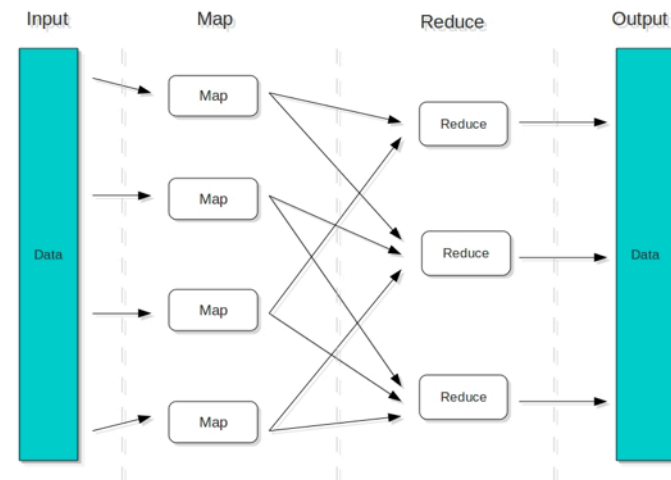
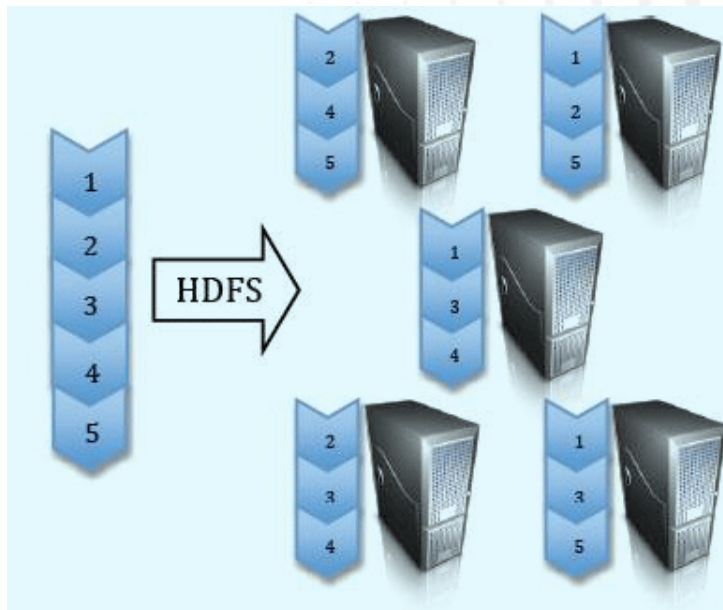


Hadoop in a Nutshell



Replicated Distributed Big-Data File System

Map-Reduce - framework for writing massively parallel jobs



Hadoop Benefits

- Reliable solution based on unreliable hardware
- Designed for large files
- **Load data first, structure later**
- Designed to maximize throughput of large scans
- Designed to maximize parallelism
- Designed to scale
- Flexible development platform
- **Solution Ecosystem**

Hadoop Limitations

- Hadoop is scalable but not fast
- Batteries not included
- Instrumentation not included either
- Well-known reliability limitations



The Oracle logo, consisting of the word "ORACLE" in white capital letters on a red rectangular background.

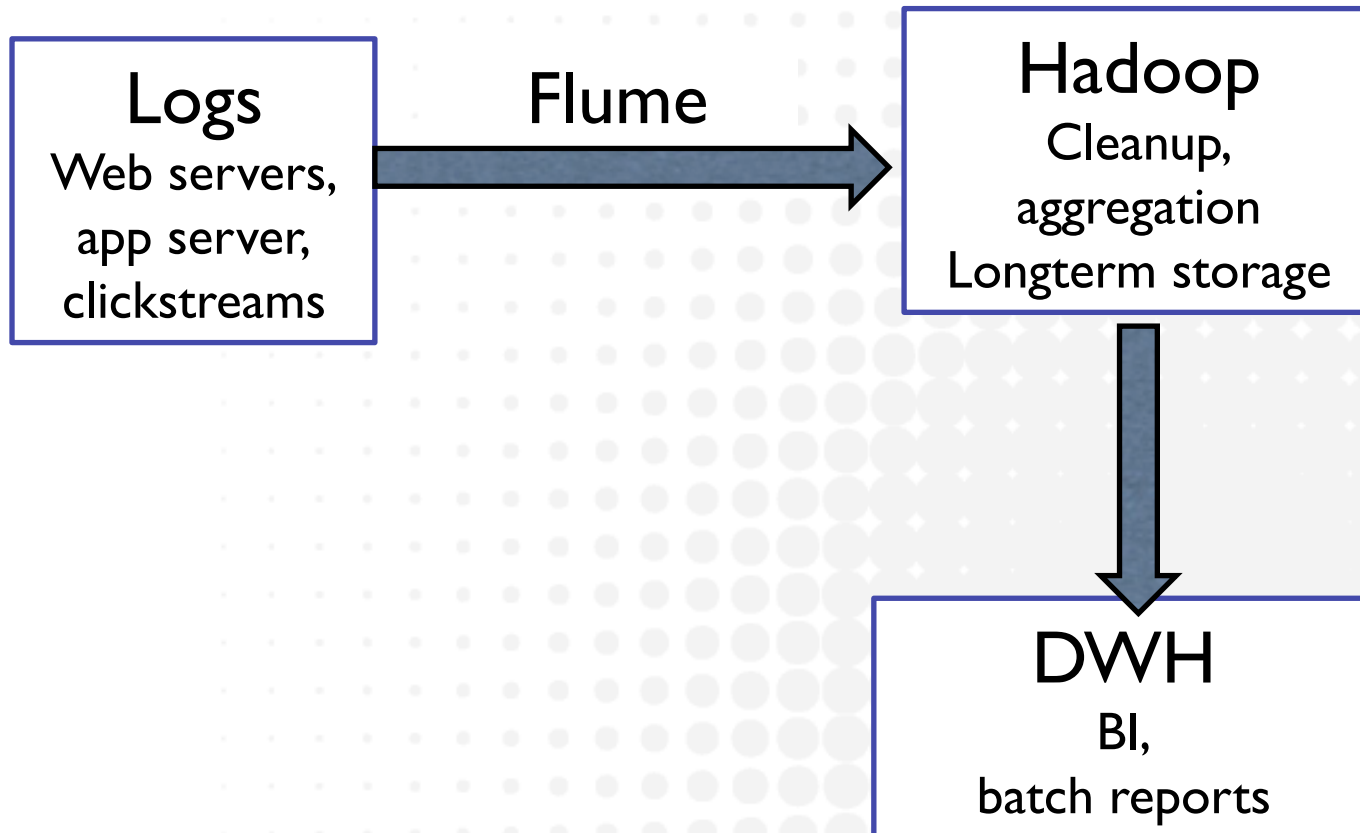
Platinum
Partner

Hadoop In the Data Warehouse

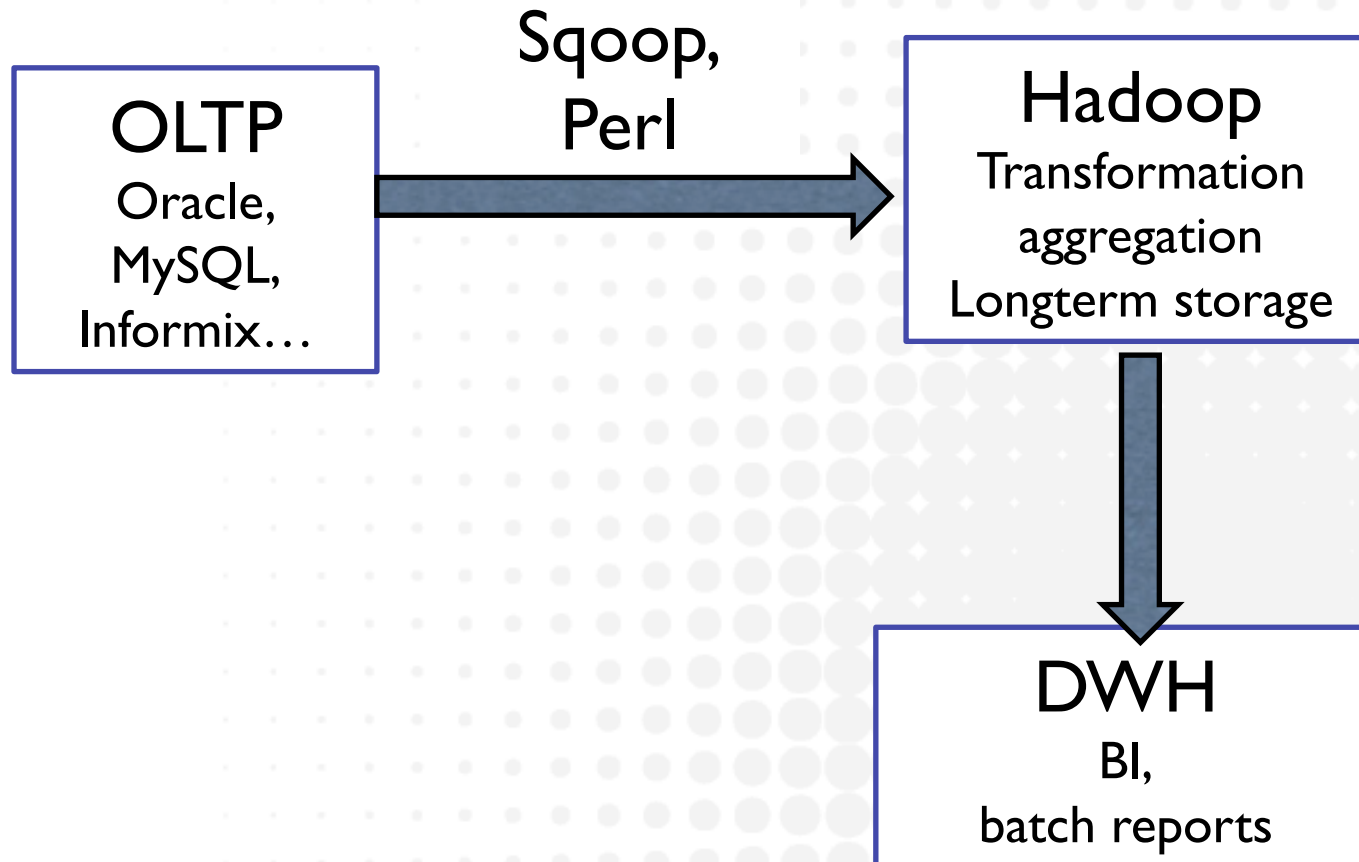
Use Cases and Customer Stories

The Pythian logo, featuring the word "Pythian" in a large, white, sans-serif font, with the tagline "love your data" in a smaller, white, sans-serif font below it. The background is a dark grey/black area with a pattern of small, light grey circles.

ETL for Unstructured Data



ETL for Structured Data



Bring the World into Your Datacenter



Rare Historical Report



Find Needle in Haystack



We are not doing SQL anymore



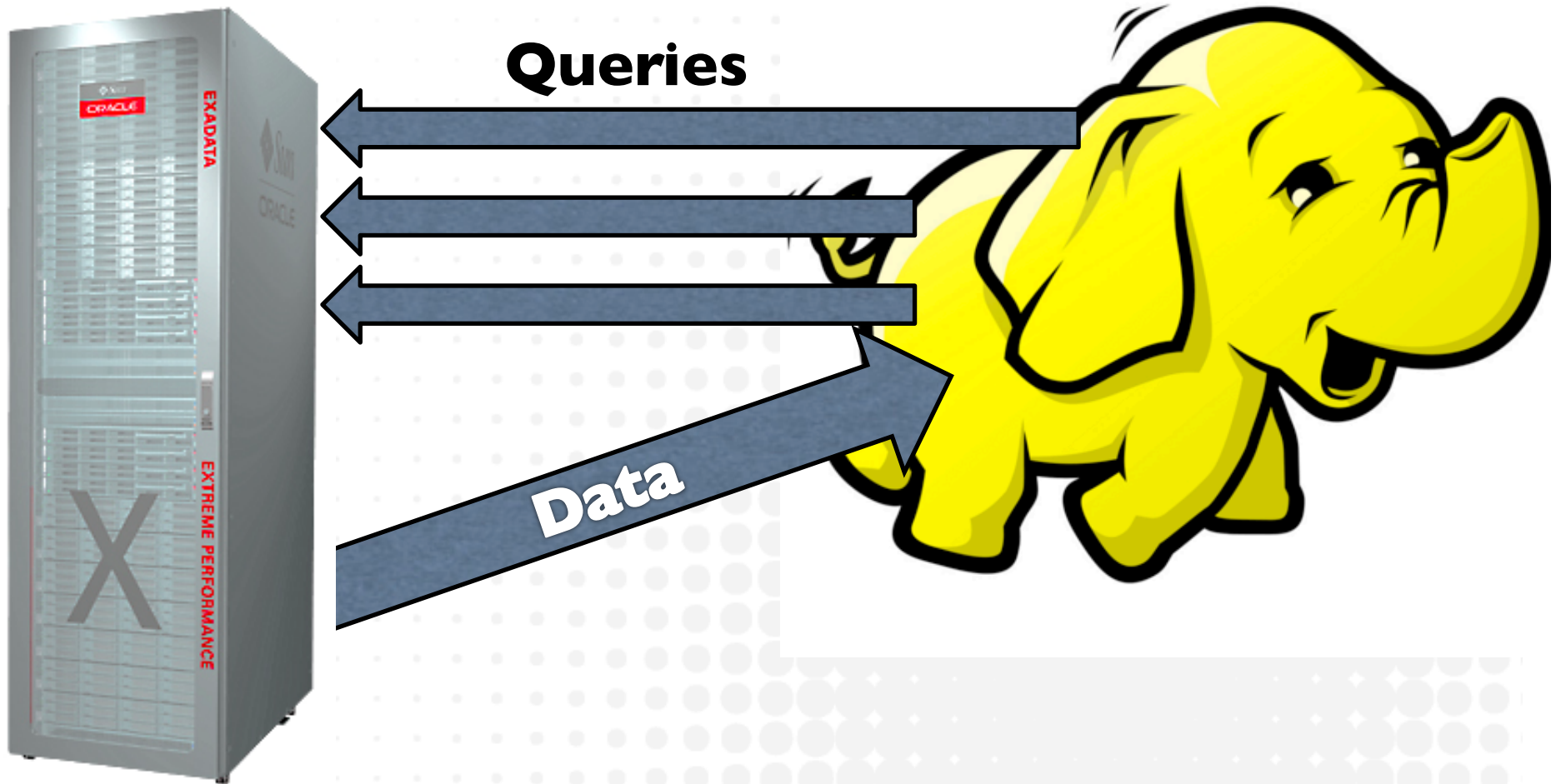
The Oracle logo, consisting of the word "ORACLE" in white, uppercase letters inside a red rectangular box.

Platinum
Partner

Connecting the (big) Dots

Pythian
love your data

Sqoop



Sqoop is Flexible Import

- Select <columns> from <table> where <condition>
- Or <write your own query>
- Split column
- Parallel
- Incremental
- File formats

Sqoop Import Examples

- Sqoop import --connect jdbc:oracle:thin:@//dbserver:1521/masterdb
--username hr --table emp
--where "start_date > '01-01-2012'"
- Sqoop import jdbc:oracle:thin:@//dbserver:1521/masterdb
--username myuser
--table shops --split-by shop_id
--num-mappers 16

Must be indexed or partitioned to avoid 16 full table scans

Less Flexible Export

- 100 row batch inserts
- Commit every 100 batches
- Parallel export
- Merge vs. Insert

Example:

```
sqoop export  
--connect jdbc:mysql://db.example.com/foo  
--table bar  
--export-dir /results/bar_data
```

FUSE-DFS

- Mount HDFS on Oracle server:
 - `sudo yum install hadoop-0.20-fuse`
 - `hadoop-fuse-dfs dfs://<name_node_hostname>:<namenode_port>
<mount_point>`
- Use external tables to load data into Oracle
- File Formats may vary
- All ETL best practices apply

Oracle Loader for Hadoop

- Load data from Hadoop into Oracle
- Map-Reduce job inside Hadoop
- Converts data types, partitions and sorts
- Direct path loads
- Reduces CPU utilization on database
- NEW:
 - Support for Avro
 - Support for compression codecs



Oracle Direct Connector to HDFS

- Create external tables of files in HDFS
- `PREPROCESSOR HDFS_BIN_PATH:hdfs_stream`
- All the features of External Tables
- Tested (by Oracle) as 5 times faster (GB/s) than FUSE-DFS

Oracle SQL Connector for HDFS

- Map-Reduce Java program
- Creates an external table
- Can use Hive Metastore for schema
- Optimized for parallel queries
- Supports Avro and compression

The Oracle logo, consisting of the word "ORACLE" in white capital letters on a red rectangular background.

Platinum
Partner

How not to Fail

Pythian
love your data

Data That Belong in RDBMS




Prepare for Migration



Use Hadoop Efficiently

- Understand your bottlenecks:
 - CPU, storage or network?
- Reduce use of temporary data:
 - All data is over the network
 - Written to disk in triplicate.
- Eliminate unbalanced workloads
- Offload work to RDBMS
- Fine-tune optimization with Map-Reduce





**Your Data
is NOT
as BIG
as you think**

Getting Started

- Pick a Business Problem
- Acquire Data
- Use right tool for the job
- Hadoop can start on the cheap
- Integrate the systems
- Analyze data
- Get operational



Thank you & Q&A

To contact us...



sales@pythian.com



1-866-PYTHIAN

To follow us...



<http://www.pythian.com/news/>



<http://www.facebook.com/pages/The-Pythian-Group/>



<http://twitter.com/pythian>



<http://www.linkedin.com/company/pythian>